Characterization and Differentiation of Three British Population Groups

PhD Thesis

Queen Mary University of London

By

David Ballard

May 2013

Blizard Institute
Bart's & The London, Queen Mary's School of Medicine and Dentistry
London, UK

Abstract

The British population is made up of three main ethnic groups: Caucasian, Afro-Caribbean and South Asian. The history of Britain is littered with a series of invasion and colonisation events, potentially resulting in a variety of different genetic influences shaping the native population. More recent immigration trends have lead to over 11 million people within the UK describing themselves as belonging to an ethnic minority. The aim of this research is therefore to characterise these three population groups for a series of genetic markers, in the process gaining an insight into the genetics and origins of the individuals within these groups and ultimately developing a robust population-of-origin classification system for a DNA sample of unknown origin.

To this end, three distinct areas of the genome were investigated. This comprised the development of a suite of PCR multiplex reactions to analyse 11 Y chromosome short tandem repeat (STR) markers, sequencing of the maternally inherited mitochondrial DNA, and analysis of a number of autosomal Single Nucleotide Polymorphisms (SNPs) known to show population specific allele distributions. The results from these studies led to the development of simple (Y chromosome and mitochondria) or complex (SNPs) classification systems enabling unknown DNA samples to be categorised into one of these 3 ethnic groups with a high degree of certainty: the Y-STR population-of-origin classification algorithm had a success rate of 80%, the mitochondrial version a 90% success rate while correct prediction was achieved over 94% of the time with the autosomal SNPs.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

A – Adenine

AMOVA – Analysis of Molecular Variance

bp – Base pairs

C- Cytosine

CI – Confidence Interval

cM - Centimorgan

DNA – Deoxyribonucleic Acid

dNTP – Deoxyribonucleotide Triphosphate

G – Guanine

HVI – Hypervariable region I

HVII - Hypervariable region II

kb - Kilobase

Mb - Megabase

MSY – Male Specific Y

NCBI – National Centre for Biotechnology Information

PCR – Polymerase Chain Reaction

rCRF – Revised Cambridge Reference Sequence

RNA – Ribonucleic Acid

SNP – Single Nucleotide Polymorphism

SRY – Sex determining region Y

STR – Short tandem repeat

T -Thymine

Acknowledgments

I would first like to thank Denise for giving me the opportunity to do this work and Barbara for initiating the project. Special mention should go to Chris Phillips for his help and advise over the years, and also to Paul for his guidance. This work would assuredly not have been possible without the help and support of everyone in the laboratory, especially Esther, Cheryl and Catherine, but also including Gary, Ekta, Kerry-Ann, Jaimin, Gabriella, Chun-Wai, Andrew, Athina and Lesley, and I am very grateful for everything they have done over the last few years. Lastly I would like to thank my family who have endured this for the past decade: my Mum and Dad, my Aunt, Angela, James, Sam, Luke, and my Granddad who had surprising faith it would eventually be finished even if he couldn't see the final product.

1 Introduction

The individuals residing in Great Britain are a diverse group with influences from many different cultures both throughout the course of history and more recently *via* modern population migration. The aim of this research is to characterise the three major ethnic populations within Britain (White, Black and Asian) and to devise methods for accurately distinguishing between these groups on the basis of DNA, the final test ideally being applicable in the determination of ethnic origin from an unknown DNA trace found at a crime scene.

1.1 The Population of The United Kingdom

The United Kingdom (UK) consists of 4 separate countries: England, Scotland, Wales and Northern Ireland. In 2004 the population of the UK was estimated to be 59.8 million people, up 19% from 1951 (50.3 million). There is no uniform distribution across the constituent countries, with England the home to 83.7% of the populace. The recent population expansion (up 3.3% in the last decade) has been driven more by the influx of foreign nationals into the UK (immigration) and less by natural increase (more births than deaths) [1].

1.1.1 Birth of the British Population

The origins of the modern British Caucasian population are numerous and convoluted. Historically there has been much gene flow into this population due to repeated small scale invasion and colonisation; although due to the geographical separation from mainland Europe, it should be noted that there have only been two complete conquests, by the Romans in 54 B.C., and the Normans in 1066 A.D.[2].

The island that now comprises Great Britain was once joined with continental Europe, and it is during this time, at least seven hundred thousand years ago [3], that early *Homo erectus* settlers first arrived. These were later superseded by *Homo sapiens*

originating from Africa [4-6], before the two landmasses split during the Mesolithic period (*circa* 7th millennium B.C.) to create the English Channel.

It wasn't until the 6th century B.C. that Britain began to receive a steady influx of Celtic people that supplanted the primitive natives, but the country couldn't appropriately be called a Celtic nation until the large population movements of the Iron Age brought great numbers of Celts across the channel [7]. Archaeological evidence suggests that the Celtic people originated from tribes in the area surrounding the upper Danube, predominately in what is now central Germany [2].

Over the next 600 years, the Celtic peoples displaced the primitive natives and divided the country up on a tribal basis, each tribe governed by its own King. Having established Roman control over Belgium and northern France in 57 B.C., Julius Caesar set his sights on Britain and commenced his first invasion in 55 B.C. He returned a year later and proceeded to conquer vast swathes of the country before returning to France, having made peace in return for an annual offering.

It wasn't until nearly 100 years later (44 A.D.) that the Roman army returned to Britain, partially at the request of the warring Celtic tribes. Emperor Claudius led four legions to subdue the populace and create Britannia, the latest province in the Roman Empire. Colonisation began soon afterwards, starting with the creation of Colchester and St Albans.

When the Roman Legions left four centuries later, the power vacuum was exploited by the invading Barbarians: the Scots coming from Ireland, the Anglo-Saxons from the European mainland and from the North came the Picts, who had previously been kept out by Hadrian's Wall. The orderly Britain forged by the Romans fragmented into many small kingdoms ruled independently and the country descended into the Dark Ages.

It is during this time that Anglo-Saxon England was forged, and once again these invaders originated from tribes in Northern Europe; specifically in the area between the mouths of the rivers Rhine and Elbe (the name England is derived from one of these tribes, the *Engle*). By the end of the 6th century A.D. virtually all of England

was under the rule of the Anglo-Saxons, with the country split into different kingdoms such as Wessex in the south-west, Mercia in the Midlands and Northumbria in the north.

By the 9th Century A.D. Britain was again under attack, this time from the Vikings. These invaders from Scandinavia repeatedly invaded, plundered and later settled the Anglo-Saxon kingdoms until only Wessex remained. Under King Alfred the Great, the Anglo-Saxons of Wessex repelled the Vikings and gradually the kingdom expanded, capturing London in 886, leaving England divided between the Anglo-Saxons in the south and the Danes in the north and east. Over the next 100 years the conquering Danes were subjugated and the country returned to Anglo-Saxon rule before a new wave of invasion by the Norsemen bought the country once more under Viking control by the beginning of the 11th century A.D.

When Edward the Confessor died in 1066, Harold ascended to the throne of England. A century earlier a Viking warrior had created the duchy of Normandy in Northern France; the Duke in 1066 was William, the warrior's grandson, later to be known as William the Conqueror. Under the leadership of William, the Normans invaded 9 months after Harold was crowned and succeeded in conquering the country, killing Harold on the battlefield at Hastings. Within a few years so complete was the subjugation that Anglo-Saxon England effectively ceased to exist, the civilisation swept away with a new ruling class, the introduction of the feudal system, a revolutionised church, a demoted status for women and even the replacement of the native Anglo-Saxon language with French and Latin in the ruling class [2]. This resulted in a destitute life for almost all the citizens, but also proved to be a success for the Normans since this was to be the final full-scale invasion and population suppression in the history of the British Isles.

Hence by the 2nd millennium, the background of the White British population was genetically diverse with influences from the Celts, Romans, Anglo-Saxons, Viking and Normans.

1.1.2 Modern Britain

Today the cultural and ethnic diversity within the UK is highlighted by the 2011 census [8]. The proportion of people in England and Wales calling themselves white British fell to 80.5% in 2011 (see Table 1.1) and in some areas of London the white British population were a minority ethnicity (more than 80% of people in the London boroughs of Newham and Brent describe themselves as from a non-white British background). The largest ethnic minorities are Black (black Caribbean 1.1%, black African 1.9%) and South Asian (2.5% Indian, 2.0% Pakistani, 0.8% Bangladeshi) individuals. Other significant minority groups are Chinese (0.7%) and mixed race (2.2%) while white non-British individuals accounted for 5.5%. These figures may be an underestimation since it has been suggested [9] that ethnic minorities are disproportionately underrepresented in a census, particularly young men in urban areas. Data from the 2011 census states that currently 13% of the population residing within the UK were born overseas (this includes 0.73% born in Ireland) [10].

Table 1.1 Self-declared ethnicity from the 2011 census for England and Wales

Ethnicity	Percentage of Population
White British	80.5
White Irish	0.9
White Other	4.6
Black	3.3
South Asian	5.3
Chinese	0.7
Other Asian	1.5
Arab	0.4
Mixed Ethnicity	2.2
Other	0.6

* White other includes a large number of individuals born in Poland, Germany, South Africa and the United States of America

When analysing the demographics of the various British ethnic populations, distinct characteristics can be seen with different groups. The Black Caribbean population structure is pyramidal in shape with a swelling at the pre-retirement age [11] (data as

of the 1991 census). This reflects the fact that immigration was actively encouraged from the Caribbean colonies after the Second World War to fill the void caused by the conflict. The broad base of the pyramid represents the recent generations from this first migratory wave. In contrast the Black African population demographic is quite different and additionally shows a more transitory residence in Britain [11], often on student visas. The predominance in the Black African population of relatively young individuals is a reflection of the young age of the large number of first generation immigrants. The rise in the Black African population between 2001-2004 is primarily caused by an increase in asylum applicants, particularly from Zimbabwe and Somalia [12], and while the proportion of individuals describing themselves as Black Caribbean stayed constant at 1.1% between the 2001 and 2011 census, the Black African proportion doubled from 0.9% to 1.8% [8].

In comparison to the changing Black African population, those with Indian ancestry comprise the longest-established and largest ethnic group within the UK, with good representation at all age levels. Immigration restrictions in the mid-1960s led to a high level of movement in both the Indian and Pakistani population *via* the further settling of dependants joining relatives already established within the UK, rather than novel immigration. This explains the unusually large number of later middle-aged individuals within these populations. Migration from Pakistan peaked in the late 1960s and early 1970s. Partly due to the same 1962 Commonwealth Immigrants Act, Bangladeshi immigration did not start in significant numbers until later than their South Asian neighbours, and didn't peak until the 1980s [11].

With 11 million people in England and Wales belonging to ethnic minorities, and 7.5 million of these born overseas, a significant proportion of the population has a different genetic background to the longer established white British population. The two most established groups are Black and South Asian individuals, and it is these two populations along with the 'native' White British population that will be studied during this research. Studies of these groups will provide population data for the investigated markers - necessary if these are to be used in identification or kinship analysis; provide a more detailed picture of the genetic background of these populations, including information on ethnic ancestry; and allow methods to be

developed that classify individuals into these three ethnic groups based on a DNA sample.

1.2 Genetic Variation

Variation between individuals can be observed in a number of different ways and take many different forms; from the obvious such as physical appearance, to subtle molecular changes resulting in alternate versions of enzymes and molecules (e.g. the cell membrane transporter for chloride ions in cystic fibrosis[13]). While some aspects of variation are influenced by the environment, many are largely determined by an individual's DNA.

DNA (deoxyribonucleic acid) is a molecule that has been likened to a blueprint for the body. It is present in virtually every cell type within the body and consists of a series of linked bases: Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). At this juncture, the function, if any, of the majority of this DNA is poorly understood. It is known that 1.2% [14] of the human eukaryotic genome is coding sequence contained within genes. At an elementary level, it is the sequence of these bases that determines the structure of a protein coded for by an individual gene. These proteins carry out many functions in the body and subtle changes between individuals in protein sequence and conformation can be important in determining such diverse factors as baldness, nose shape and disease susceptibility. It is additionally known that some other sections of the DNA regulate gene expression through interaction with transcription factors and it is suggested that some non-coding DNA is functionally important once transcribed into RNA ([15] & [16]).

In each cell, rather than one long DNA molecule, the DNA is separated into 46 smaller segments, called chromosomes. Of these 46 chromosomes, two determine sex; these are the X and Y chromosomes. The other 44 chromosomes, known as the autosomal chromosomes, can be grouped into 22 pairs, with one of each pair being inherited from the mother and one from the father. The fact that an offspring has half its DNA from each parent ensures a child is different from the parents and facilitates evolution of the species.

Any two unrelated humans of the same sex will show about 99.9% concordance in their DNA sequences [17], however the 0.1% differences are vitally important. These differences can take many forms, from major chromosomal anomalies to changes at a single nucleotide position. Three of the most common sources of variation are copy number variance, length polymorphisms (e.g. short tandem repeats (STRs)) and Single Nucleotide Polymorphisms (SNPs).

1.2.1 Copy Number Variation

Copy number variations (CNVs) collectively encompass deletions, insertions and complex multi-site variants [18]. The size of the sequence involved in these CNVs range from kilobases (kb) to megabases (Mb), but a CNV has been defined as a DNA segment at least 1kb in length displaying a variation in copy number within a population compared with a reference genome [18]. The genetic variation conferred by these CNVs can result in many functional changes within an organism compared with other individuals in the population, among other things this can result in altered disease susceptibility or indeed be the cause of some disorders.

1.2.2 Length Polymorphism

Length polymorphism relates to the insertion or deletion (indels) of nucleotides at a specific point in the DNA. The most frequently occurring forms of indel polymorphism involve only a few bases [19] and occur at a disproportionately high frequency within areas of the genome containing DNA sequence that are repeated multiple times [19]. It is estimated that nearly 50% [19] of the human genome is made up from such repetitive elements of DNA. The type of sequence being repeated determines the classification given to the stretch of repetitive DNA.

Microsatellite markers, also known as short tandem repeats (STRs), have short repeat motifs (usually 1-6 bp) that are duplicated in series. It is estimated that there are over a million microsatellite loci within the human genome (depending on the definition used) [20], accounting for 7% of chromosomal DNA [19]. Variation between individuals can occur in the actual number of repeated segments at any one STR loci

(see Figure 1.1). This is exploited in the criminal justice field to identify individuals. In the UK, 10 STR markers are interrogated across the genome using a commercially available kit called SGM+ (Applied Biosystems) and the probability of two individuals possessing the same 10 marker SGM+ genotype is about 1 in 10^{12} in a US Caucasian population [21], demonstrating how polymorphic these markers can be.

Figure 1.1 Example of STR repeat structure

Variation/mutation in repeat number at microsatellite loci is believed to occur during DNA replication through a process called polymerase slippage [22]. This is a phenomenon whereby the DNA polymerase dissociates from the DNA strand during replication of an STR and re-attaches in the wrong place, hence changing the number of repeat units at that specific short tandem repeat locus (see Figure 1.2). It has been suggested that this model is too simplistic to account for the wide variety in microsatellite mutation rate, and other factors implicated include a role for DNA repair mechanisms [23], allele length [24], and the contextual sequence of the locus [25]. Due to the wide range of variables that can have an influence on mutability, there is no uniform STR mutation rate, but large studies tend to arrive at an average overall rate in the order of 1×10^{-3} mutations per locus per generation [26, 27]. Over 85% of all mutations involve the gain or loss of 1 repeat unit [24].

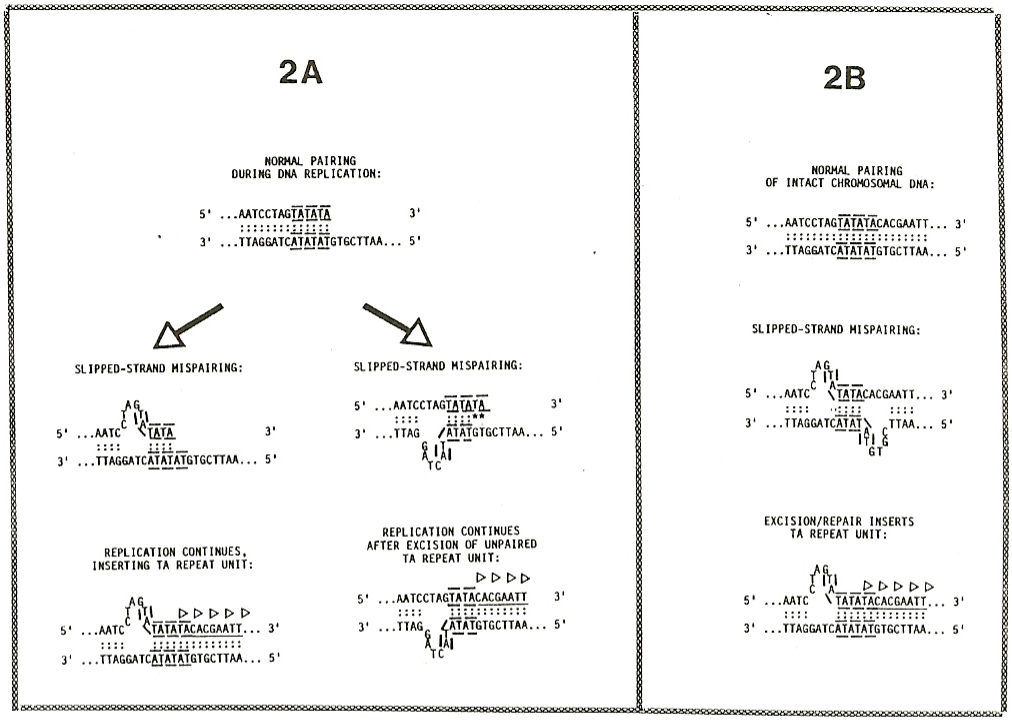


FIG. 2.—Generation of duplications or deletions by SSM between contiguous repeats. Small arrows indicate direction and starting point of DNA synthesis; colons indicate base pairing. A, 2-Base slippage in an AT-repeat during replication of a DNA duplex, followed by continued chain elongation. Slippage in the 3′ → 5′ direction (left panel) results in insertion of one AT unit; slippage in the other direction (right panel) results in deletion of one repeat unit. The deletion shown on the right results from excision of the unpaired repeat unit (asterisks) at the 3′ end of the growing strand, presumably by the 3′ → 5′ exonuclease activity of DNA polymerase. B, The same slip occurring in intact duplex DNA. Mismatched regions form single-stranded loops, which may be targets for excision and repair. Results depend on where excision/repair events take place: excision of the shorter loop on the top strand, followed by repair synthesis using the lower strand as template, results in addition of one AT repeat unit as shown; other outcomes, including deletions, are also possible.

Figure 1.2 A description of polymerase slippage.
Taken from Levinson & Gutman [22]

1.2.3 Single Nucleotide Polymorphisms

The most important form of DNA variation in terms of sheer number of polymorphic sites is Single Nucleotide Polymorphisms (SNPs). A SNP is defined as a single nucleotide position in a genome that possesses at least two alleles within a population: the minor allele being present at an appreciable frequency, traditionally at least 1%.

The publication of the first diploid sequence from a single individual [19] shows that SNPs represent 26% of all base variances when comparing this sequence to the human reference sequence curated by the NCBI (National Centre for Biotechnology Information, a division within the US National Institutes of Health). However this figure accounts for the substantial total of 78% of all DNA variants observed since

other events such as insertions can involve multiple bases. The HapMap project have estimated that there are approximately 10 million SNPs in the human genome where the minimum allele is present at a frequency of at least 1% across the worldwide population [17]. A recent international collaborative project to re-sequence 1092 complete genomes from 14 European, sub-Saharan African, American and East Asian populations produced a data set containing 38 million SNPs (where the minor allele was seen in at least 1 individual), which it is estimated captures up to 98% of common (>1%) SNP variants in associated populations [28].

SNPs usually express only 2 alleles within a population, most often either C/T or A/G. Adenine and Guanine are bases composed of a double ring purine structure while Cytosine and Thymine are nucleotide bases with a single ring pyrimidine structure. Transition from an A:T nucleotide to a G:C nucleotide has been shown to occur nearly three times more frequently than a transversion event (a change between types of base structure, e.g. from a purine A to a pyrimidine C) while the reverse transitions from G:C to A:T occur over 5 times more frequently than transversions [29]. These mutation events occur at a far lower rate for SNPs than microsatellite markers (average mammalian point mutation rates are in the region of 2.2×10^{-9} per base pair per year [30]), however the local rates are known to vary widely, with a two fold change in substitution rate across the genome and many mutational hotspots [31].

1.3 Differentiating Populations

When Charles Darwin published 'On the Origin of the Species' in the nineteenth century, this expounded the theory of evolution - in the main concerned with the adaptation, and birth of, an entire species over eons. Diversification on a smaller scale, and over a shorter time period, is evident in the differences between sub-populations of a species, and only some of these are due to selective pressures. In humans this is visually apparent in a number of ways, from skin tone and body size, to facial characteristics and hair colour.

1.3.1 Population Genetics

Population genetics is the field of biology dedicated to studying the variation in allele frequency between populations of the same species that occurs over time. There are four main mechanisms by which the allele frequency distribution can change within a population:

Mutation – A germ-line mutation is a change in the DNA between a parent and child. This can either give rise to a totally new allele within the population, or increase the frequency of an existing allele at the expense of the allele possessed by the parent that would normally have been transferred to the child. Different types of genetic marker mutate at different rates, as explained in section 1.2.

Selection – The process by which an allele increases in frequency within a population because it confers an advantageous characteristic to the organism, e.g. an increased chance of surviving to a reproductive age or an increased attractiveness to the opposite sex. The converse is also true: alleles associated with deleterious effects on the individual decrease in frequency over generations.

Migration – The introduction into the population of individuals from another separately evolved population.

Genetic drift – The tendency of allele frequencies to alter within a population due to the stochastic pattern of mating (some genetic lines die out while others spawn many offspring) [32]. The size of the population has a large impact on the magnitude with which genetic drift can influence allele distribution (assuming random mating, the smaller the population the greater the likely effect). There are two important types of event that can occur throughout a population's development, both increasing the influence of genetic drift by reducing the population size. These are:

> Bottleneck – An event occurring in history (e.g. a natural disaster) that wipes out a large proportion of the population, hence the descendants arise from a smaller gene pool and allele frequencies are different from neighbouring populations that comprise a more varied genetic background. In the Middle

Ages it has been suggested that the Black Death may have produced a bottleneck effect reducing the genetic diversity in the British population [33].

Founder effect – This has a similar outcome to a bottleneck, but arises when a small subset of a larger population becomes separated, e.g. by colonising an island.

These four evolutionary processes have acted upon *Homo sapiens* whilst the species has evolved and expanded creating genetic differences between populations. These differences are predominately along geographical lines, although there are also cultural divides – for example hierarchy (e.g. the Indian Hindu caste system [34]) or religion (e.g. the Ashkenazi Jews [35, 36]). The first publication of such genetic variation within human populations occurred back in 1919 and subsequently altered the understanding of the biology behind the ABO blood grouping system [37].

Below, the concept of genetic differentiation is expanded, giving examples of different areas of the human genome that have been shown to vary between populations.

1.3.2 Pigmentation Genes

A clear way to distinguish population groups would be to use the genetics underlying a visually defining feature of the population. The most obvious candidate characteristic for this approach on a broad continental level is probably skin pigmentation. The drawback, as with the genetics underlying other physical traits, is that the phenotype observed is not determined by one specific allele at one DNA locus, but is the result of complex interplay between a myriad of changes at numerous genes. In mice, 127 genes are known to affect pigmentation, and 68 of these have human homologues [38].

Skin pigmentation in indigenous human populations is correlated to the levels of ultraviolet radiation in the autumn [39]. The key to colour, whether it be in skin, hair or the eyes, is a compound called melanin. The precise hue achieved is known to be influenced by many factors including the composition and production of the melanin,

the packaging into melanosomes and the distribution of these melanosomes from melanocytes to various cells. A large genetic association study on 6 candidate genes has confirmed that while similar genetics underlie the dark skin in West African and Island Melanesian populations, lighter skin in East Asian and European populations evolved independently through mutations in at least 3 genes: SLC24A5 (regulates melanosomal calcium level), TYR (produces tyrosinase, a critical enzyme in the early stages of melanin synthesis) and MATP (important in tyrosinase transport and intracellular processing) [40]. Additionally, mutations in the melanocyte stimulating hormone receptor (MC1R) are disproportionately associated with individuals displaying red hair/fair skin phenotypes [41] and worldwide pigmentation variance is known to be affected by at least 2 other genes: ASIP (agouti-signalling protein that binds to MC1R) and OCA2 (a melanosomal membrane transporter) [40]. Other areas of the genome recently shown to influence human pigmentation include 6p25.3, SLC24A4 and KITLG [38].

1.3.3 Non-coding DNA

As already stated, most DNA does not code for proteins, and is not therefore highly conserved between individuals. Variations in these areas of the genome have already been shown to differentiate groups of people by association rather than a direct cause (i.e. a specific variant is associated just with one population due to demographic factors rather than a coding variant in a gene that could directly differentiate them, as discussed in section 1.3.2).

Detailed below are some of the areas of genetic variation that can be exploited to separate DNA samples into distinct groups based upon their population of origin.

1.3.3.1 Copy Number Variants

While copy number variants (CNVs) are not exclusively related to non-coding DNA, there is no evidence linking them directly with the phenotypic traits distinct between ethnic groups, hence any association between CNVs and ethnicity is likely due to demographic history (e.g. bottlenecks). Reproduced below in Figure 1.3 are the results obtained when 67 biallelic CNVs were genotyped for 210 samples [18]. A

model-based Bayesian clustering method was employed to infer population structure from the data and assign all samples to one or more populations *via* a probabilistic determination [42]. Optimum results are obtained assuming 3 ancestral populations, and it can be seen that the samples separate nicely into 3 groups using just these 67 biallelic CNVs– European (individuals from Utah with European descent (CEU)), African (Nigerian samples from the Yoruba tribe (YRI)) and East Asian (Japanese samples from Tokyo (JPT) and Han Chinese samples from Beijing (CHB)).

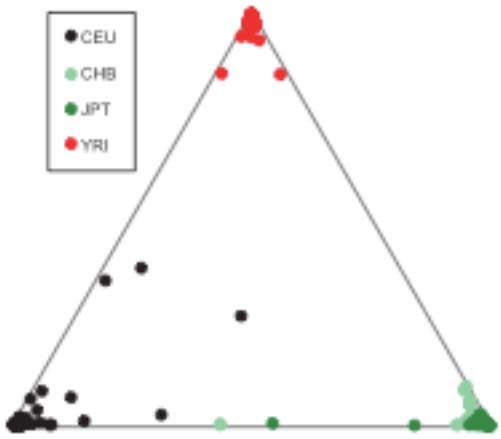


Figure 1.3 CNV population differentiation
Data from Redon *et al.* [18] displaying a graphical representation of 210 samples from 4 populations separated into 3 groups by a statistical clustering method utilizing the results obtained from 67 biallelic CNVs. Each corner of the equilateral triangle represents one of the 3 assumed ancestral population groups, and the nearer a sample is to an apex of the triangle, the better the fit into that group. It can be observed that nearly all the samples cluster very nicely into 3 groups (European, African, East Asian) based on the genotypes of the 67 CNV markers.

1.3.3.2 Microsatellite Markers

The CEPH Human Genome Diversity Cell Line Panel is commercially available and consists of 1056 samples from 52 defined populations representing indigenous peoples from across the world. In 2002, Rosenberg *et al.* [43] published a major study focussed on genotyping the CEPH panel for 377 microsatellite markers. In line with other estimates, this study showed that most human variation is explained by the differences between unrelated individuals within population groups and only a small proportion of worldwide human variance is as the result of additional differences between populations. Reanalysis of the data by Excoffier & Hamilton [44], using a stepwise mutation model incorporating the possibility of recurrent mutations, gave figures for these different sources of variance, showing that 81-85% of variation occurs between individuals within populations, 4-6% between populations within

regions (e.g. France and Germany in Europe), and 10-13% between major regions (e.g. Europe and Africa).

By identifying the differences in allele distribution across all 377 loci, a model based clustering algorithm [42] was able to separate out samples into groups containing individuals with similar genotypes. Samples invariably clustered into geographically distinct areas despite the fact that no information relating to sample origin was provided to the model. Based solely on the genotypes from these 377 STR markers, when the model assumes the data consists of 5 distinct population groups (k=5), the resulting clusters correspond to sub-Saharan Africa, Oceania, the Americas, East Asian and Eurasia (Europe, the Middle East, and South/Central Asia). This is represented graphically in Figure 1.4 Some further refinement within continental groups is also possible, as shown in Figure 1.5.

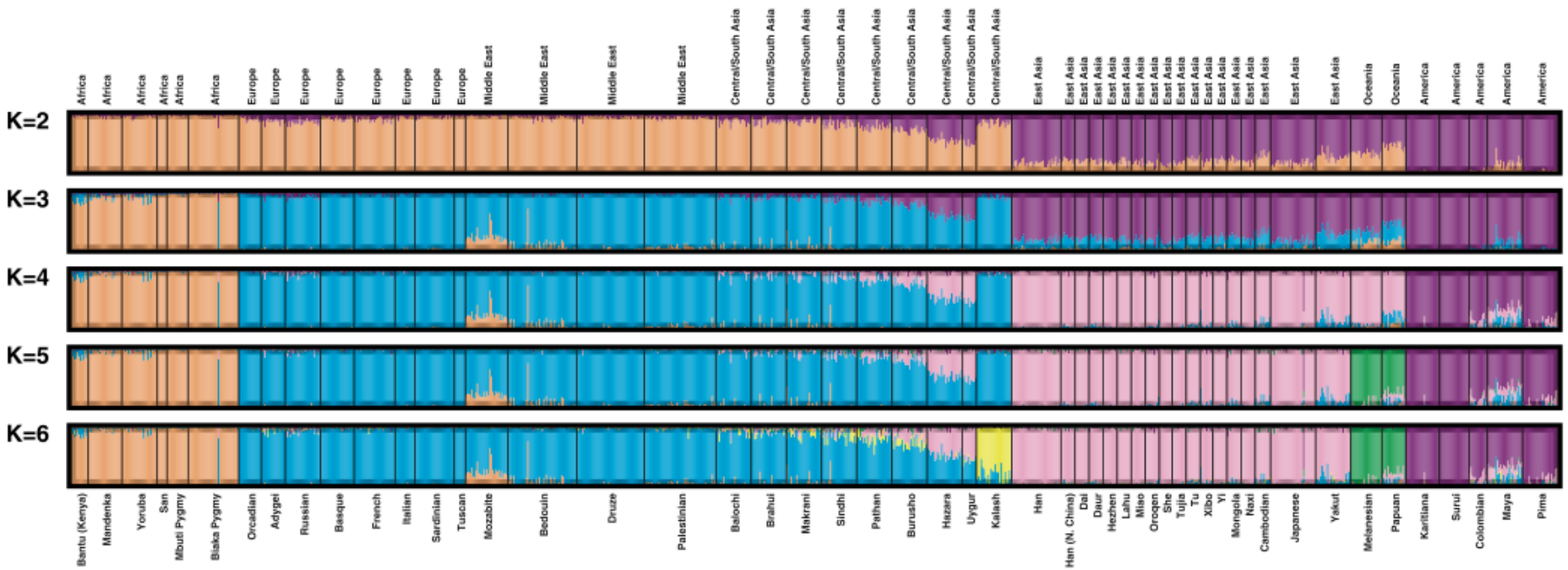


Figure 1.4 Clustering of 1056 individuals from 52 native worldwide populations using 377 STRs
Populations are separated by dark lines, and each sample within a population is represented by a thin coloured bar – the distribution of colour within this bar is dependent on the affinity of the sample to the different clusters generated by the model. The model assumes the data fits into k clusters. Populations are labelled below, and continent designations are given above such that the arrangement from left to right is in the order Africa, Europe, Middle East, Central/South Asia, East Asian, Oceania and America. Figure from Rosenberg *et al.* [43].

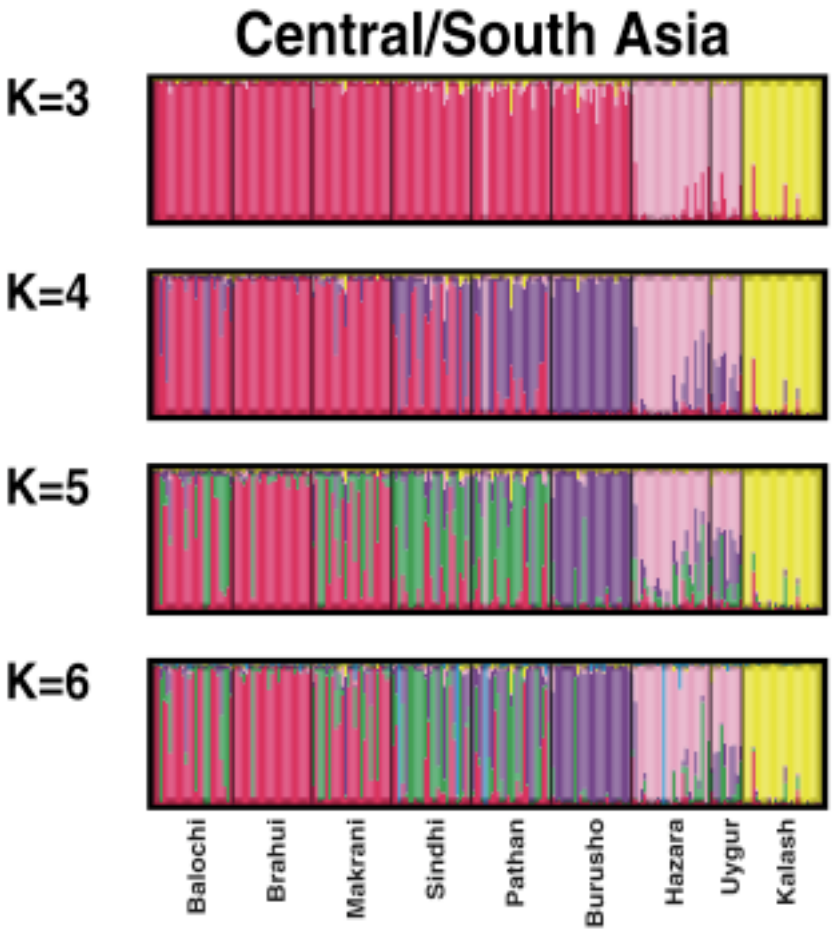


Figure 1.5 Clustering within the Central/South Asian populations using 377 STRs
Populations are separated by dark lines, and each sample within a population is represented by a thin coloured bar – the distribution of colour within this bar is dependent on the affinity of the sample to the different clusters generated by the model. The model assumes the data fits into k clusters. Genetically, culturally and physically (blue eyes, light skin) distinct from the rest of Central Asia, the Kalash tribe is from very high up in the Hindu Kush mountain range of northwest Pakistan. The Burusho population is a linguistically distinct population from north Pakistan, hence why it may tend to separate from the other Pakistani samples. It is suggested that shared Mongol ancestry may cause the Hazara (Pakistan) and Uygur (northwest China) populations to cluster together. Figure from Rosenberg *et al.* [43]

An attempt has also been made to differentiate populations using just 6 of the polymorphic STR markers that are tested in the UK during routine forensic profiling [43]. Results obtained in this study from a 5 population classification gave correct assignment in 56% of Caucasian samples, 67% of Afro-Caribbean sample, 43% of Indian sub-continental samples, 66% of Southeast Asian samples and 30% of Middle Eastern samples. Figure 1.6 graphically displays the classification of samples from an independent study examining 3 major population groups with 20 STR markers [45]. The poor results obtained from both studies highlight the disadvantages in using a limited set of the highly mutable and polymorphic microsatellite markers for population discrimination applications. Conversely, in fields such as forensic

genetics, analysis often needs to be performed on samples compromised in both quality and quantity making the analysis of numerous STR loci impossible.

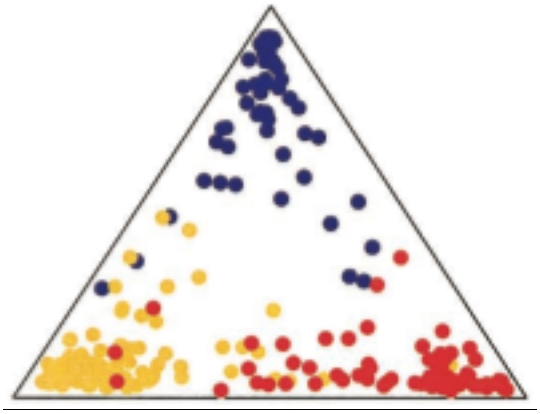


Figure 1.6 Population Separation Using 20 STRs
Dots represent individuals from East Asia (Red), sub-Saharan Africa (Blue) and Europe (Yellow). Genotypes from 20 STR loci are used for each individual to determine the proportion of ancestry from 3 computer-generated populations. The results are represented graphically in the same format as Figure 1.3 – the nearer to a point of the triangle that an individual is, the better the sample fits into that statistically derived population group. The spread of dots within the triangle demonstrates that clustering with only 20 STRs is not good even when examining 3 well differentiated populations. Figure from Bamshad *et al.* [45].

1.3.3.3 *Alu* Insertions

Shorts Interspersed Nuclear Elements (SINES) are repeated sequences spread throughout the genome. In total there are about 1.5 million of these sequences, each measuring 100-300 bp in length [46]. Of these, the *Alu* motif is the most common, accounting for about two thirds of all SINES [46]. *Alu* elements are genotyped as either being present or absent from a particular location within the human genome. They are stable once inserted into the genome, so can be used to study evolution back to primates as well as modern human population stratification.

Figure 1.7 shows the results obtained when 100 *Alu* elements were typed across 23 ethnic groups from Europe, East Asia and sub-Saharan Africa [45]. Analysis of the results with *Structure* [42] assigned each sample to 1 of 3 groups (shown to broadly refer to the 3 continents sampled) while the 95% confidence intervals were estimated with a bootstrap technique. These results demonstrate that while good classification outcomes can be achieved with relatively few *Alu* markers, many more are needed before a success rate approaching 100% is achieved, especially when separating the more closely related European and Asian populations.

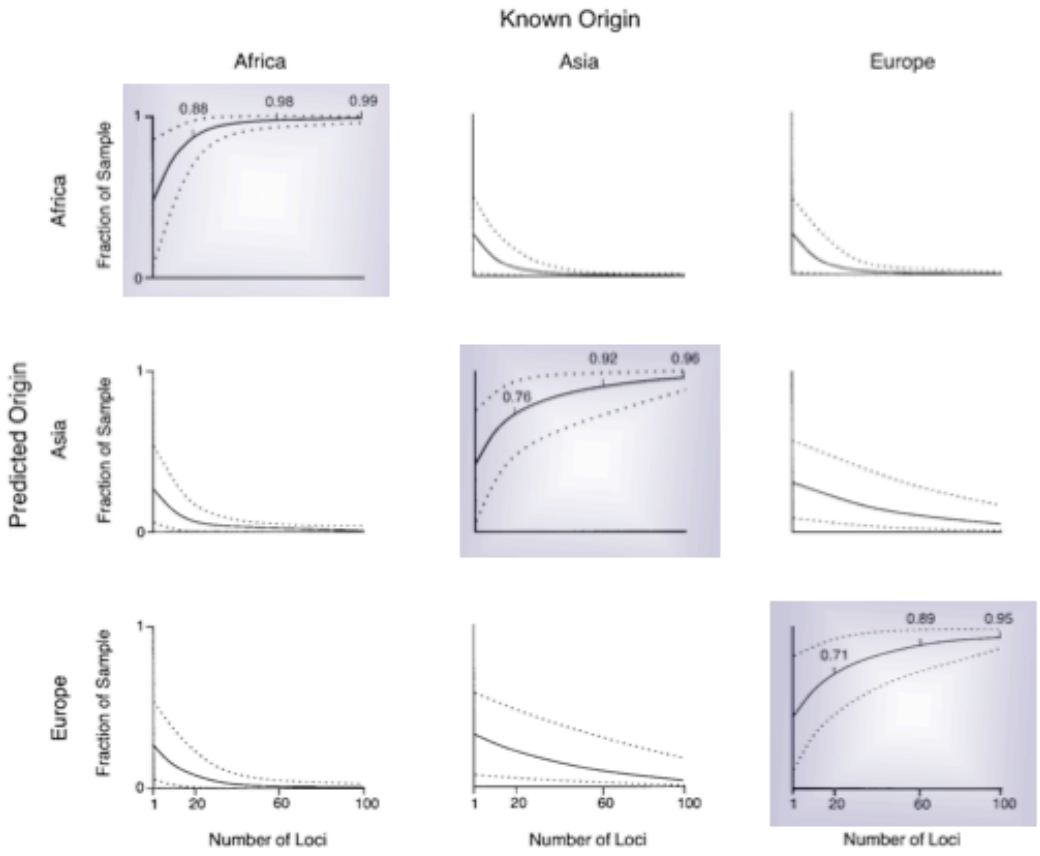


Figure 1.7 Population classification using *Alu* markers
A set of graphs showing predicted origin versus known origin for 3 sets of samples. Individual graphs detail the fraction of samples classified into that group (with 95% confidence intervals) against the number of *Alu* markers used for the classification (from 1-100). Highlighted graphs represent the correctly classified samples. Adapted from Bamshad *et al.* [45].

1.3.3.4 SNPs

A large-scale SNP study examining 1,586,383 SNPs in Americans of European, African and East Asian ancestry [47] demonstrated that a sub-set of the SNP markers showed major allele frequency differences between the 3 populations. In 18% of the SNPs studied, polymorphism (observing both SNP alleles) was restricted to just one of the populations - i.e. the SNP was population specific. Subsequent to this research, Lao et al [48] analysed nearly 8,500 SNPs in 6 world population groups, and identified the 10 most informative SNPs for identifying genetic ancestry. The 51 populations of the CEPH panel were then genotyped for these 10 SNPs and fairly good population clustering was observed when specifying that the data fit into 4 different groups (relating to (1) Native America, (2) sub-Saharan Africa, (3) Oceania and East Asia plus 2 Central/South Asian populations, and (4) Europe, the Middle

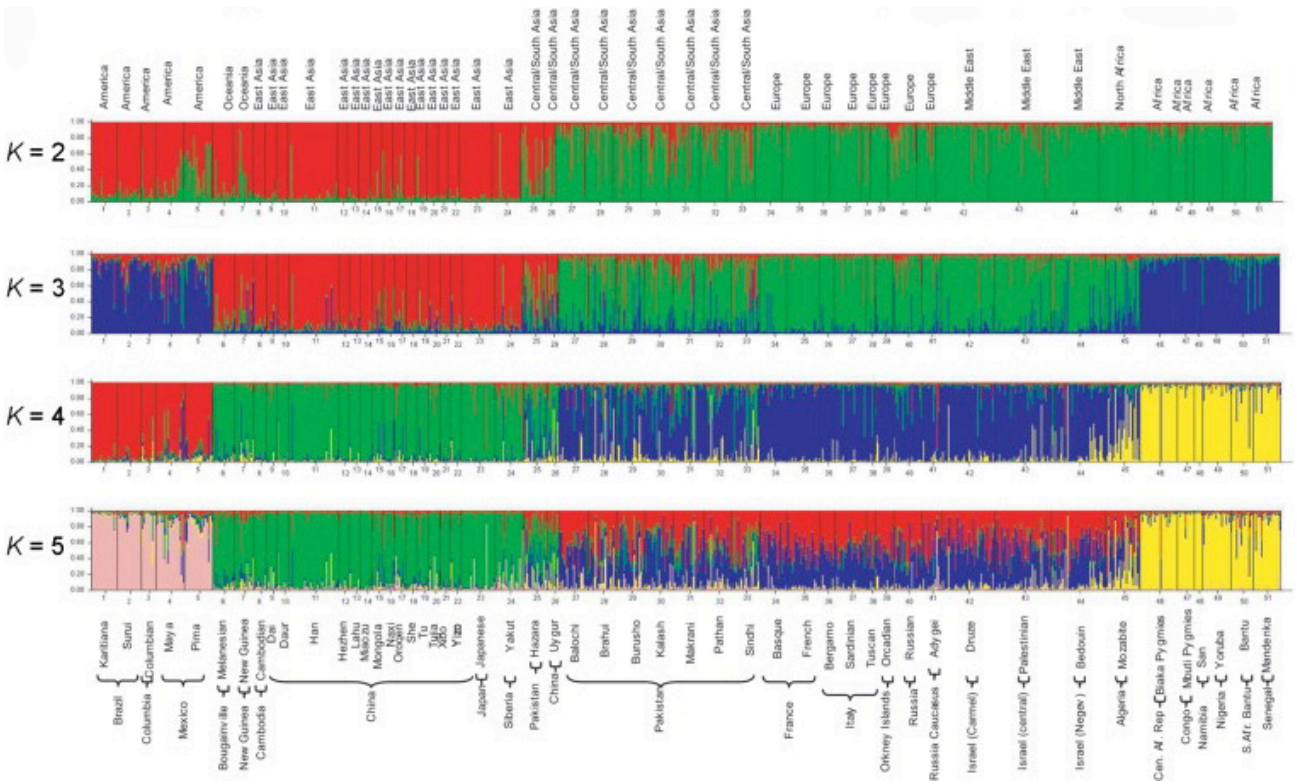


Figure 1.8 Results of a *Structure* [40] analysis on the genotypes obtained from 10 SNP markers in the 51 CEPH panel populations
Each line represents one of the 1046 samples, with the 51 populations named below and separated by a dark line. The distribution of colour within each individual sample represents the ancestral contributions of the different genetic groups. The software repeats the analysis with different numbers of ancestral groups specified, represented by the value of K. Group characteristics are determined by the software on the basis of the genotypes and not influenced by prior knowledge of population membership. Figure from Lao *et al.* [48].

East, North Africa and the remaining Central/South Asian populations); see Figure 1.8. This demonstrates that a small set of well chosen SNPs are able to produce good results when separating populations on a major continental level.

More subtle variations can also be elucidated using SNPs. A recent large scale study (5,700 SNP markers and 1094 participants) [35] on European populations has detected a degree of population structuring across the continent; statistical models showing that the European samples can be clustered genetically into 2 defined populations. The samples within these groups represent a North / South divide, with 84/86 Italians and 66/74 Spanish samples being represented in the Southern group while the majority of individuals with Western, Eastern, Central and Scandinavian European ancestry clustered together in a second, 'North' European, group. Similar clustering results were obtained when analysing only the 400 most informative SNPs of those tested.

1.3.4 Haploid Markers

Humans mostly possess a diploid genome, that is, two different copies of every chromosome, one from each parent. There are however two regions of the genome only possessed by an individual in a single, haploid, copy: the circular mitochondrial DNA passed down from mother to child, and the Y chromosome which is only transferred between father and son.

The direct transmission of a single set of linked markers down either the male or female lineages make these ideal candidates for understanding the population movements of men and women throughout history. Marked inter-population differences in allele frequency distributions of Y chromosome STRs and SNPs have already been demonstrated [49, 50]. Similar variation across the globe can be observed in mitochondrial DNA sequences [51, 52].

1.3.4.1 Y chromosome

1.3.4.1.1 Evolution

The two sex chromosomes in humans, the X and the Y, are believed to have originally evolved from an ordinary pair of autosomes about 300 million years ago [53, 54]. In mammals an XX genotype is generally found in the female of the species while an XY genotype is found in the male, although there can be exceptions (e.g. lemmings [55]). The gene present on the Y chromosome that triggers male sex determination is the SRY (sex determining region Y) which functionally stimulates testes development [56, 57]. This has an X chromosome homologue called SOX3 (SRY-related HMG box containing 3) and it was suggested by Foster and Graves [58] that both SOX3 and SRY were originally different alleles of the same SOX gene that was functionally important in the development of both male and female embryos. They hypothesised that SRY evolved from SOX3 in both function and sequence to acquire a male-specific role in testis development, either initiating autosomal chromosome divergence into sex chromosomes or occurring as a consequence of previous X-Y differentiation caused by a different sex determining gene now supplanted by SRY.

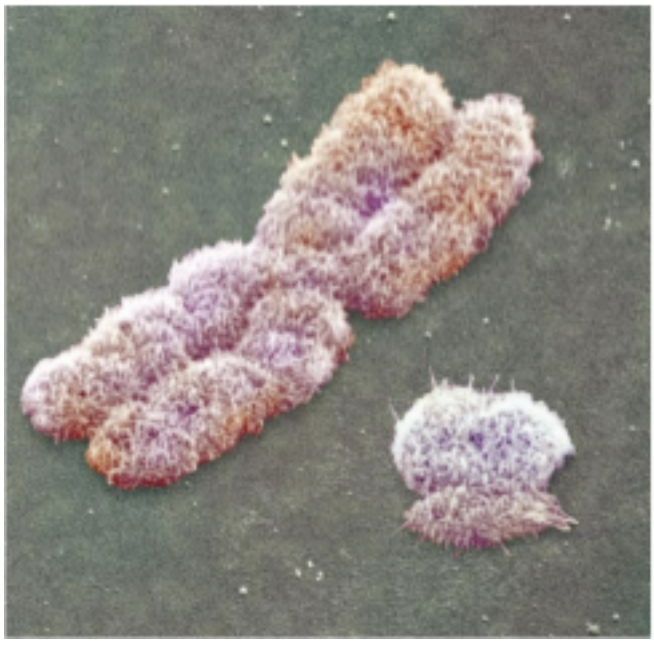


Figure 1.9 The X and Y chromosome magnified 10,000x [59]

The present-day Y chromosome spans roughly 60 million base pairs (Mb) in length, containing approximately 23Mb of euchromatatic DNA and a variable amount of heterochromatic (highly repetitive) DNA [54]. Over 95% of the Y chromosome is

male specific, having further differentiated from the X chromosome through the last 300 millennia. Lane & Page [53] postulated that this differentiation occurred in four distinct stages and it is thus suggested that X-Y recombination was suppressed one stratum (or region) at a time. These 4 strata are located sequentially on the X chromosome and recombination would appear to have been prevented by inversions of the homologous area in the Y chromosome [53]. This suppression of recombination facilitated divergence of these once similar autosomes into distinct sex chromosomes. The theory has undergone minor refinement with the publication of the Y chromosome euchromatic sequence [54] suggesting that the boundaries between some strata may be a little more blurred. Additionally comparative genomics has shown that some of these evolutionary strata are formed by the addition of autosomal DNA segments to the sex chromosomes at different evolutionary stages [60]. There are two areas of the Y chromosome comprising approximately 3Mb that still do recombine with the X chromosome during sperm production, these are at the tips of the short [61] and long [62] arm and are called pseudoautosomal regions.

Despite accounting for approximately 1% of the human diploid genome, the male-specific region of the Y chromosome encodes only 27 distinct proteins/protein families [54]. There are three main classes of sequence comprised within the male specific Y chromosome: X-transposed, X-degenerate and ampliconic, while one quarter of the euchromatic DNA is additionally shown to be arranged in eight palindromic sequences [54, 63]. The combination of sequences similar to the X chromosome (99% similarity for the X-transposed and 60-96% for the X-degenerate), combined with a high level of duplication within the Y chromosome itself, originally made sequencing and study of this chromosome very difficult; by the end of 1996 there were fewer than 60 known polymorphisms on the Y chromosome [64]. These sequence characteristics also made SNP validation problematic since many candidate SNPs were in actuality artefacts caused by incorrect sequence alignment [65]: comparing a genuine Y chromosome sequence either with a similar sequence on the X chromosome or within the Y chromosome itself (e.g. comparing palindromes which exhibit on average 99.97% identity [63]).

By 2003 over 200 SNPs had been well characterized in the Y chromosome and ordered into a hierarchical haplogroup tree dependent on the evolutionary age of the

SNP [65]. As already stated, over 95% of the Y chromosome avoids meiotic recombination (from here on referred to as MSY – Male Specific Y), and this portion is therefore passed on unchanged from father to son down the generations, barring mutations. From a large-scale re-sequencing study based on two members of the same family separated by 13 generations, it was calculated that roughly one single base change occurs each generation across the length of the euchromatic MSY (3×10^{-8} mutations/nucleotide/generation) [66]. As the mutation rate for SNPs is low, in most cases a change in nucleotide at a specific base occurs only once in evolution and is not subsequently mutated back. The linkage of all the markers in the MSY combined with this low mutation rate therefore means that Y-SNPs form forks in human evolution like branches on a tree, and all subsequent mutations within a population occurs onto the same genetic foundation. An example of a tree containing Y-SNPs defining some lineages within Africa, Asia and Oceania is given in Figure 1.10. The branches of these trees are given names (called haplogroups), and in the Y chromosome tree the major limbs are labelled according to the letters of the alphabet. Figure 1.11 shows the relative distribution of these Y chromosome haplogroups throughout the world.

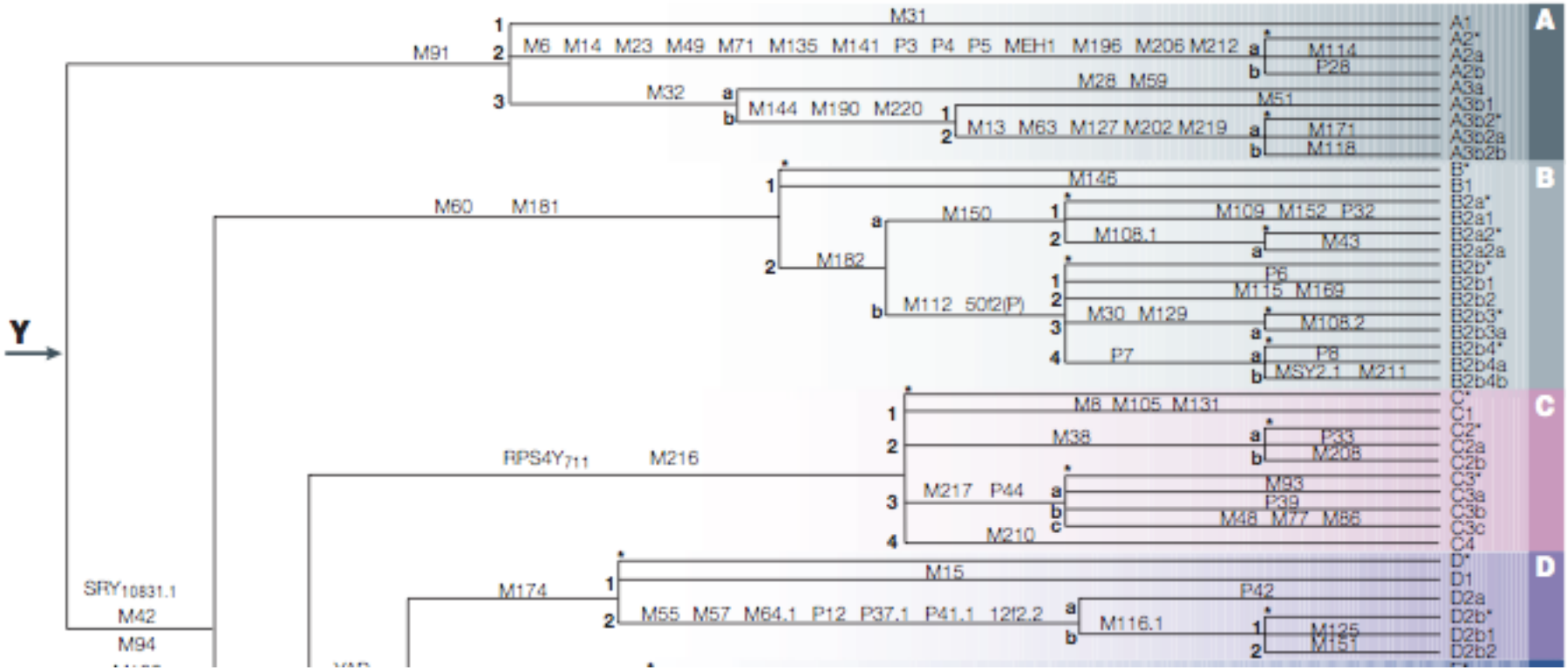


Figure 1.10 A phylogenetic tree showing the relationships between and within haplogroups A-D.
Each line represents a branch of the tree, with the markers named above it representing the diagnostic SNP or SNPs for that particular branch. The rate at which base substitutions occur is so low that in most cases it can be assumed that the change is unique in evolution, and hence if an individual possessed the mutant allele at locus M51 (and therefore belongs to haplogroup A3b1) then it is known that the only other base changes the individual will possess (for all the SNPs making up the tree) are those that had occurred in earlier ancestors at M144, M190, M220, M32 and M91. Taken from Jobling *et al.* [65].

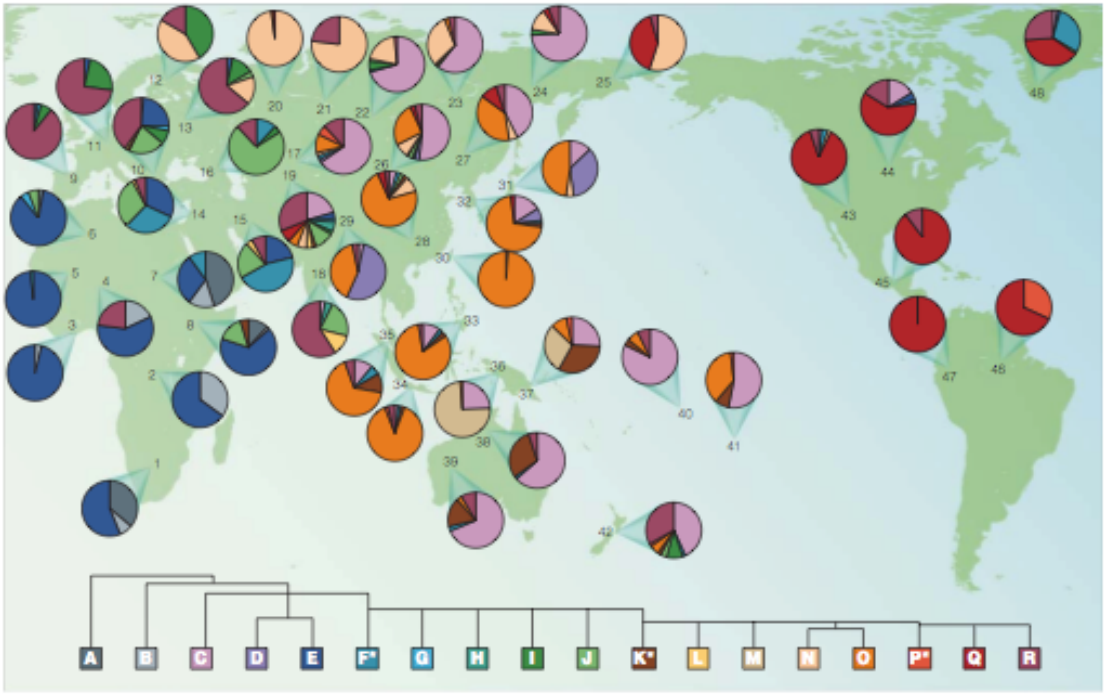


Figure 1.11 Y chromosome haplogroup distribution in 48 worldwide populations.
Figure from Jobling *et al.* [65].

Attempts have been made to use this haplogrouping system to determine the ethnic origin of an unknown DNA sample. Brion et al [67] typed over 1000 samples for 29 Y-SNPs, defining 31 major haplogroups across the world. The results showed geographic haplogroup stratification, with 16 of the 26 haplogroups being continent specific for East Asian, European and African populations. A limitation with this methodology is in the analysis of admixed populations: for example analysis of the white Brazilian population [68] suggests most Y chromosomes have a European signature while the mitochondrial DNA haplogroups are characteristic of African and Amerindian ancestry – a finding consistent with historical data showing that mating occurred between European males and African or Native American females.

1.3.4.1.2 Y Chromosome STRs

Y chromosome STRs are more sensitive in detecting differences between closely related populations because of the higher mutation rate compared with Y-SNPs. The increased incidence of mutation accentuates differences between recently diverged populations, but also increases the chance of recurrent mutations causing haplotypes

to be identical by state rather than descent [69], i.e. observing the same haplotype in individuals with different ancestry.

Y chromosome STRs have to be analysed as a haplotype due to the linked nature of all the markers. A haplotype is defined as the allelic states of a set of markers on the same chromosomes that are normally analysed together because of linkage disequilibrium due to the absence of recombination, e.g. because they are very close together on the same chromosome (a haplotype block) or because they are on a non-recombining chromosome (i.e. the Y). Since this linkage means the Y-STRs are not independent, the frequency of a haplotype cannot be calculated from the frequencies of each constituent allele contained within it, but instead must be calculated by counting the number of times that the entire combination of alleles (i.e. haplotype) is observed within a population. Hence in Y-STR analysis it is of paramount importance to have adequate haplotype databases for the populations being studied.

In 1997, Kayser *et al.* [70] characterized the 13 polymorphic Y chromosome STRs that were currently being investigated around the world. Of these 13, 7 were recommended for use as a basic set of STRs in forensic and paternity situations. These were DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392 and DYS393. These recommendations were broadly adopted and databases were established containing these 'minimal haplotype' loci (e.g. [71]).

It was further recommended [70], that the highly polymorphic bi-local loci DYS385, YCAII and YCAIII be included if a higher discrimination between unrelated males was needed. It was suggested that by adding these loci to the 7 STR set, that the discriminatory capacity could increase from a level between approximately 74-90% up to nearly 100%. All 3 of these STRs are very polymorphic due to the fact that there are two copies of each locus (bi-local) on the Y chromosome (see Figure 1.12 for DYS385 structure), hence each of the 3 STRs can produce two different peaks. The loci YCAII and YCAII both consist of dinucleotide repeat motifs, making them difficult to analyse with challenging DNA, however DYS385 is a tetranucleotide repeat and a report by Caglia *et al*. [72] illustrated that by adding just this loci to the 7 basic STRs, the individualization could be raised from 70% to 93.6%.

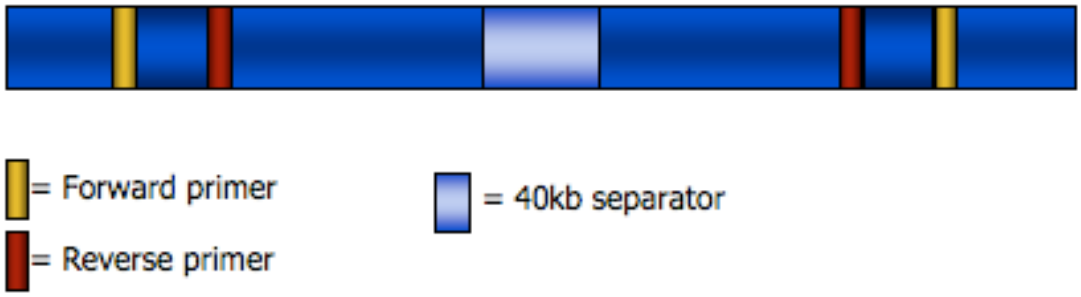


Figure 1.12 Structure of DYS385.
A 190kb fragment of DNA has been inversely duplicated with 40kb separating the copies. Since the DYS385 marker is within this 190kb section, there are 2 copies of this locus on the Y chromosome (designated 'a' and 'b'), each of which can mutate independently. This arrangement is represented graphically.

A robust pentaplex reaction comprising 5 of the recommended loci (DYS19, DYS389I, DYS389II, DYS390 and DYS393) has been developed [73] which simplified the use of Y chromosome STRs. At the time this research was commenced, two studies published on the Y chromosome had just resulted in a set of 12 novel polymorphic STR loci being discovered [74, 75]. The completion of the human genome project has since increased the number of available loci to over 200 [76], however due to the established nature of the Y-STR databases, international guidelines (Scientific Working Group on DNA Analysis Methods (SWGDAM), Y-Haplotype Reference Database (YHRD) [77]) recommend the use of the original 7 minimal haplotype loci, plus DYS385 and 2 of the newer loci discovered by Ayub *et al.* [74].

1.3.4.2 Mitochondrial DNA

1.3.4.2.1 Structure, Function and Evolution

Mitochondria are distinct organelles within the cell cytoplasm that play a crucial role in cellular respiration. They have a diameter of roughly 0.068µm [78] and are bound by a double membrane, of which the inner membrane extends inwards in folds. Mitochondria have their own DNA content independent of the cell's nuclear DNA, comprising of approximately 16,569 base pairs arranged in a circular double stranded configuration [79]. The first mitochondrial sequence was published in 1981 by Anderson *et al.* [79], and despite the significant technological challenges this project presented at the time, re-sequencing 18 years later established that there was an error rate of only 0.07% [80]. The original sequence from 1981 was labelled the Anderson

sequence, or Cambridge Reference Sequence, and was used as the reference to which all other mitochondrial sequences were compared. The subsequent re-analysis in 1999 resulted in 11 minor changes to this reference sequence, and all human mitochondrial genomes are now compared with this Revised Cambridge Reference Sequence (rCRS) [80]. Due to an uneven distribution of pyrimidine (single ring structure) and purine (double ring structure) bases between the two strands, one strand is referred to as the heavy strand (purine rich) while the other strand is referred to as the light strand (this contains an excess of pyrimidine bases).

The mitochondrial genome codes for the 12S and 16S ribosomal RNAs, 22 transfer RNAs and 13 proteins [79]. In contrast to the nuclear genome, there are few, if any, non-coding bases between the genes [79] making this a very gene-rich sequence. There is only one area with a notable collection of non-coding sequence (spanning 1.2kb), and this is designated the control region while the remaining 15.3kb of the mitochondrial genome is labelled as the coding region (see Figure 1.13).

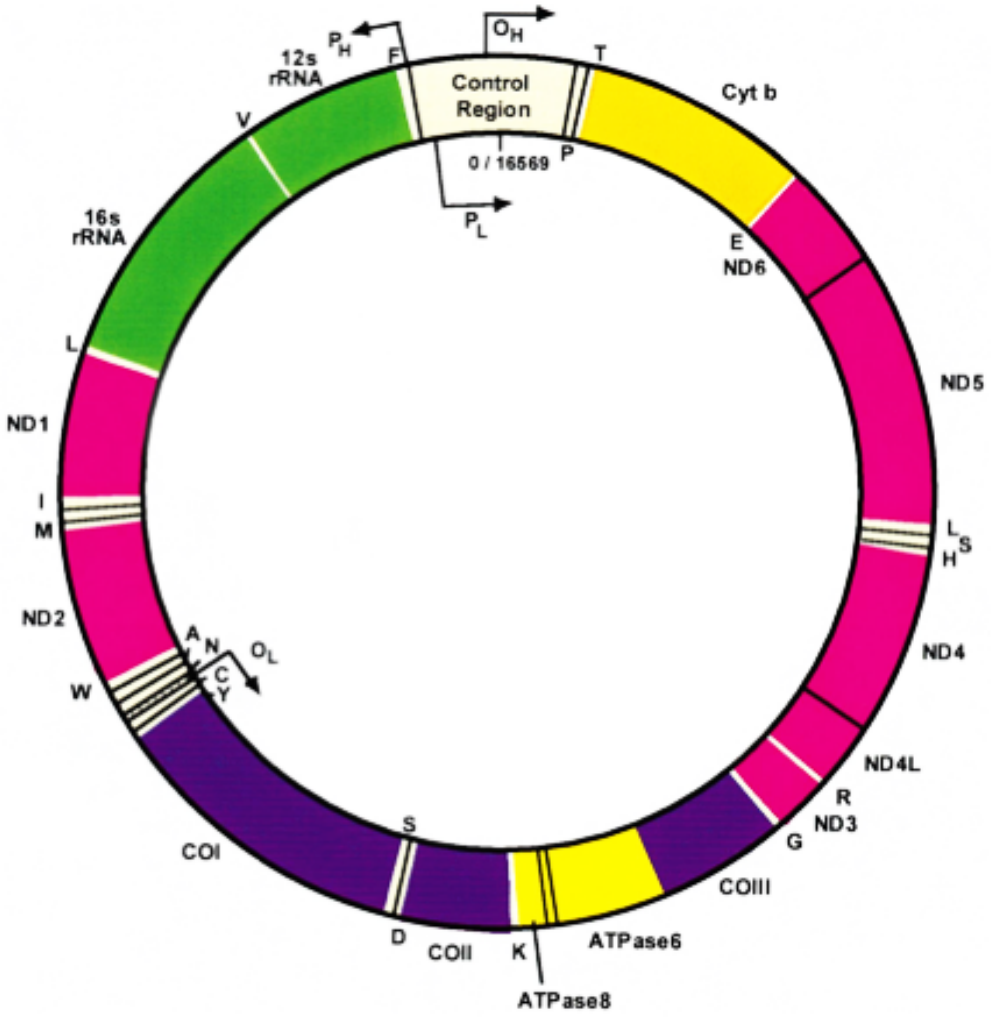


Figure 1.13 Diagrammatic representation of the mitochondrial genome
The circular, double stranded, mitochondrial genome, with the heavy strand on the outside and the light strand on the inside. The 22 transfer RNA are shown in light yellow, the 2 ribosomal RNAs in green, the genes coding for 7 subunits of Complex I are displayed in pink while the gene coding for 1 unit of Complex III is shown in orange and the three subunits of Complex IV and two units of Complex V are respectively shown in purple and deep yellow. The non-coding control region is shown at the top. Figure modified from [81].

The mitochondrial organelles provide various functions to the cell including buffering cytosolic calcium and regulating apoptosis *via* the mitochondrial permeability pore [82]. More importantly they have a major role in the process of turning glucose and oxygen into energy, a task vitally important for the cell. The mitochondria is the site of this aerobic respiration, taking the pyruvate products from glycolysis and generating energy, in the form of ATP, *via* the tricarboxylic acid cycle and the oxidative phosphorylation pathway. Additionally, the ß-oxidation of fatty acids occurs entirely within the mitochondria, ultimately also generating ATP *via* the oxidative phosphorylation pathway. All thirteen proteins coded for by the

mitochondrial DNA are involved in this oxidative phosphorylation, and changes in the DNA sequence that alter these protein are implicated in various metabolic and degenerative disorders as well as aging [82]. Various data suggests that some geographically defined sequence differences in these genes may be the result of climate adaptation, and these adaptive changes may now influence our predisposition to certain diseases [83].

It is believed that the present day mitochondria organelle developed as the result of a prokaryotic organism being assimilated and living within the cell in symbiosis. Many different strands of evidence support this theory, including the circular form of the mitochondrial genome, the lack of a nucleus and the bacterial nature of the ribosomal RNAs [84]. Experimental data supporting this theory was presented nearly 30 years ago [85], and more recent DNA sequencing results also substantiate these conclusions, more specifically pointing to an ancestor of the α-proteobacteria as being the original prokaryote [86].

Unlike nuclear DNA, which is generally present in cells in just two copies (one copy of each chromosome from each parent), the copy number of mitochondrial DNA is much higher. In 1991 Satoh and Kuroiwa [87] experimentally determined that each mitochondria within a cell contained 1-15 copies of the mitochondrial genome, at an average of 4-5 copies. Additionally each cell contains multiple mitochondria, estimated by Robin and Wong in 1998 [88] to be between roughly 80-680 depending on cell type. Other studies have since refined these figures, so while there is an average copy number of 200-1700 [88] mitochondrial genomes per cell, this can vary between 50-75 copies in a spermatozoon [89] up to the order of 100,000 in a maturing primary oocyte [90]. The presence of multiple copies of the mitochondrial genome helps to explain why it can often be analysed in situations where nuclear DNA has degraded (e.g. ancient bones). Furthermore, the cellular location of the mitochondrial DNA (in the mitochondria organelles) also has a strikingly protective function in regard to DNA degradation [84].

Inheritance of this mitochondrial DNA has been an area beset with controversy and widespread misconceptions, however it is a fact that during fertilisation the tail of the sperm does enter the oocyte, and hence mitochondrial DNA is introduced from the

paternal gamete into the newly formed zygote [91, 92]. It has been suggested that poor resolution with some molecular biology techniques may be behind the fact that only the maternal mitochondrial sequence is observed in the child [91] (i.e. the paternal mitochondrial DNA is present but is outnumbered to such an extent that is can't be observed), however it appears most likely that this is not the case and that an active cellular process occurs to remove the paternal mitochondrial DNA from the zygote [93, 94]. It is even the case that paternal mitochondrial genomes are lacking in children conceived with the assisted pregnancy technique intracytoplasmic sperm injection when the entire spermatozoon (including mitochondria) is injected into the oocyte [89]. There have also been reports of paternal mitochondrial recombination (e.g. [95-97]), however most of these have since been shown to be of dubious provenance [98], and it is now widely accepted that there is no routine paternal mitochondrial DNA contribution during human reproduction. Mitochondrial DNA is therefore inherited directly from mother to child, and will be identical between all maternally related individuals barring mutations.

The mitochondrial genome is known to replicate during cell division [99], however replication also occurs in post-mitotic cells. It's been established that there is a constant turnover of mitochondria in post-mitotic cells, and in mice the mitochondrial half-life has been calculated to be 8-23 days depending on cell type [100]. The synthesis phase of replication is believed to take only about 75 minutes [100], however there is some ambiguity as to the precise mechanism by which this replication takes place [99, 101-103]. Two competing theories state that replication occurs either bi-directionally from multiple origins downstream of bp191 (the previously defined origin-of-replication O_H) [102], or that two separate unidirectional replication events occur, the heavy strand initiating at O_H while synthesis of the lighter strand can begin at multiple locations, including the previously defined O_L [103].

As mitochondrial DNA doesn't recombine (maternal inheritance), mutation events are the only method by which the DNA sequence can be altered. Mutagenic reactive oxygen species (such as the superoxide free radical or hydrogen peroxide) are generated by mitochondria during the production of ATP; superoxide being produced in complexes I and III of the electron transport chain [104]. The rate of reactive

oxygen species production is tissue dependent, but it is estimated that up to 0.15% of electron flow during mitochondrial respiration could be eventually converted into hydrogen peroxide [104]. Reactive oxygen species can attack proteins and lipids, but it was hypothesized in 1972 that the free radicals generated within the matrix of the mitochondria could directly interact with the mitochondrial DNA, resulting in the accumulation of DNA damage [105]. This damage manifests as various DNA lesions, including abasic sites and modified bases or sugar residues [106, 107]. Pyrimidine bases are more susceptible to oxidative damage than purine bases [108].

Mitochondrial mutations can also occur *via* chemical attack. Lipophilic carcinogens such as polyacrylic aromatic hydrocarbons, azo dyes and nitrosoamines accumulate within the mitochondrial membranes resulting in DNA damage [106]. The overall rate of mutations in mitochondrial DNA is found to be about 10 fold higher than in nuclear DNA [109]. Studies on the control region have calculated this rate to be about 1 mutation/base/million years (Myr) using pedigree analysis [110] whilst phylogenetic analysis generally estimates a slightly lower rate, for example 0.087 mutations/bp/Myr [111].

Given that there are multiple copies of mitochondrial DNA within each cell, a mutation in one copy will lead to heteroplasmy – more than one version of the mitochondrial sequence within the cell. If mitochondrial mutations occur in somatic cells, then over time it is possible for the mutated version of this sequence to reach a high frequency within individual cells, most likely by random genetic drift (a process whereby the random nature of mitochondrial DNA replication and turnover alters the relative frequency of the mutant and original sequences within cells while random segregation of mitochondrial during cell division can further alter the balance) [112, 113]. Clonal expansion of this cell over time can lead to localized accumulation of cells containing this mitochondrial mutation [114]. It is suggested that in post-mitotic cells, such as those in the CNS, only somatic mutations occurring in childhood or early adulthood are biologically significant due to the number of years that it takes for genetic drift to accrue the mutation at a high frequency [113]. The accumulation of mitochondrial deletions with age is one method that is being investigated as a tool for age determination when presented with an unknown DNA sample [115].

Given how these mutations have been shown to accumulate, the very high number of mitochondrial DNA copies within oocytes (orders of magnitude more), and the relatively low number of cell divisions needed to produce mature oocytes, it would be expected that a germline mutation would show up as heteroplasmy in the child and future generations – the presence of both the maternal and mutant sequences. In fact this type of point heteroplasmy (heteroplasmy at a single base) is quite rare and segregation between generations occurs quite rapidly following a mutation so that either the mutant base becomes fixed homoplasmically within the population or disappears altogether [90, 116, 117]. It has been determined that this mitochondrial DNA segregation (resulting in a rapid change in the ratio of the heteroplasmic variants between generations) occurs very early during development as oogonia differentiate into primary oocytes [90]. This is important because, unlike mature oocytes, oogonia still have only a small pool of mitochondria [118], and hence a large effect can be exerted by genetic drift both during the stochastic mitochondrial replication occurring continuously within the cells and the segregation of mitochondria into daughter cells during the mitotic development process [90]. Whether this genetic drift is aided by a deliberate reduction in mitochondrial DNA copy number in primordial germ cells is still a contentious topic [116, 119-121].

Whilst it has been calculated in mice that it would take an average of 15 generations for a germline mutation to become fixed homoplasmically within a maternal lineage [90], empirical data from maternal relatives suggests that this often occurs far more rapidly, and in fact cases have been reported where complete allele switching essentially occurs in a single generation in both cattle [122] and humans [117, 123]. Recent statistical analysis comparing published work on mouse models with human clinical data has suggested that this observation of rapid human transitions highlights a species specific phenomenon: there are greater levels of heteroplasmy variance between generations in humans than in mice [124].

Intraovarian selection also ensures that deleterious mutations created within the gene-coding region of the mitochondrial DNA are actively selected against before ovulation (and hence transmission to a child) [82]. This means that while mitochondrial lineages can be characterized by generally non-harmful changes within the coding region, the non-coding control region proportionally expresses much more

variation. Within this control region there are specific regions showing heightened variability (hypervariable regions 1-3) [125] and furthermore the individual mutation rate of nucleotides can also vary [126], with some being characterized as hyper-mutable, possibly due to the sequence context [127].

While it has been stated that point heteroplasmy is observed fairly rarely, length heteroplasmy is a feature often present within mitochondrial DNA genomes. Length heteroplasmy occurs when there are multiple versions of the mitochondrial genome, each version differing due to the insertion or deletion of a base. This phenomenon is often observed in the poly-C stretches within the control region (between bases 16184-16193 and 303-315). In these cases the different versions of the genome vary in the number of consecutive cytosine nucleotides present, e.g. an individual may contain DNA with 9, 10 or 11 C nucleotide runs after 16184. Replication slippage is posited to be the cause of this length variation [128], which can occur at the level of the tissue, cell or mitochondria organelle [129]. Length heteroplasmy is more likely to be observed in individuals with long C tract repeats; for example when investigating the C stretch between nucleotides 303-310 Lutz-Bonengel *et al*. [130] found that only 5% of individuals that were homoplasmic at this location had a repeat stretch of 8 or more Cs, while 84% of individuals heteroplasmic at this region exhibited 8 or more Cs as their major sequence – furthermore as 61% of those in the study were heteroplasmic the long C nucleotide runs were the more prevalent type within the sample cohort. Levels of length heteroplasmy can differ between studies depending on the sensitivity threshold for detecting the minor sequence(s), and the population investigated, but a large collaborative study of over 5000 individuals [131] found that over half contained some form of length heteroplasmy within the control region.

Given the strict non-recombining maternal inheritance of mitochondrial DNA, and the ability of mutations to persist and alter the germline sequence despite the high mitochondrial copy number, mutations build-up sequentially during history in much the same way as on the Y chromosome (i.e. when new mutations arise in a lineage, they do so on the background of all other mutations that have occurred in that lineage during recent evolution). Analysis of mitochondrial sequences from across the globe has led to the construction of a complex phylogenetic tree that details all the major

enduring lineages that have arisen throughout recent human evolution, and how they relate to each other [132]. Due to the relative high mitochondrial mutation rate, there are numerous SNPs in this tree that are shown to mutate independently multiple times throughout history, and hence be present in more than one lineage, in contrast to the more stable situation with Y chromosome SNPs. New minor branching sections are routinely being added to this phylogenetic tree as more full mitochondrial sequences are typed.

1.3.4.2.2 Worldwide genetic structure and influence on mitochondrial DNA distribution

Various archaeological studies have demonstrated that modern humans evolved in Africa around 200,000 years ago, well before Neanderthals had vanished from Eurasia [133, 134]. Subsequent expansion across the globe occurred in the last 100,000 years [135] in 2 distinct waves. One proposed route is east from Africa along the coast of the Indian Ocean, populating Oceana and Southeast Asia about 60,000 years ago [136]. It has been calculated that this large-scale colonization event was the result of just a few hundred women emigrating from Africa [137]. The second, more recent, event populating western Eurasia occurred around 45,000 years ago either directly from Africa *via* the Sinai Peninsular and Levant (Eastern Mediterranean) [138, 139] or from an early offshoot of the emigration event 20,000 years earlier (made possible once the climate had improved sufficiently to allow passage into the Levant and Europe) [137]. Much of the data to support these claims has come from mitochondrial DNA analysis, and is explained by the mitochondrial DNA distribution across the globe.

There are currently more than 2000 ethno-linguistic groups originating from Africa, representing 30% of the world's languages and highlighting the diversity within the continent [140]. This is reflected in the mitochondrial DNA variation [52] where greater diversity is seen within Africa than in any other continent [141]. A study in 2009 [142] on autosomal genetic variation within Africa established that the continent could be split geographically into 6 different genetic groups (see Figure 1.14). *Structure* analysis of this data could further identify more refined substructuring. A number of African populations exhibited a low degree of European/Middle Eastern

geneflow, including Cape Coloured individuals from South Africa and Eastern Afroasiatic-speakers.

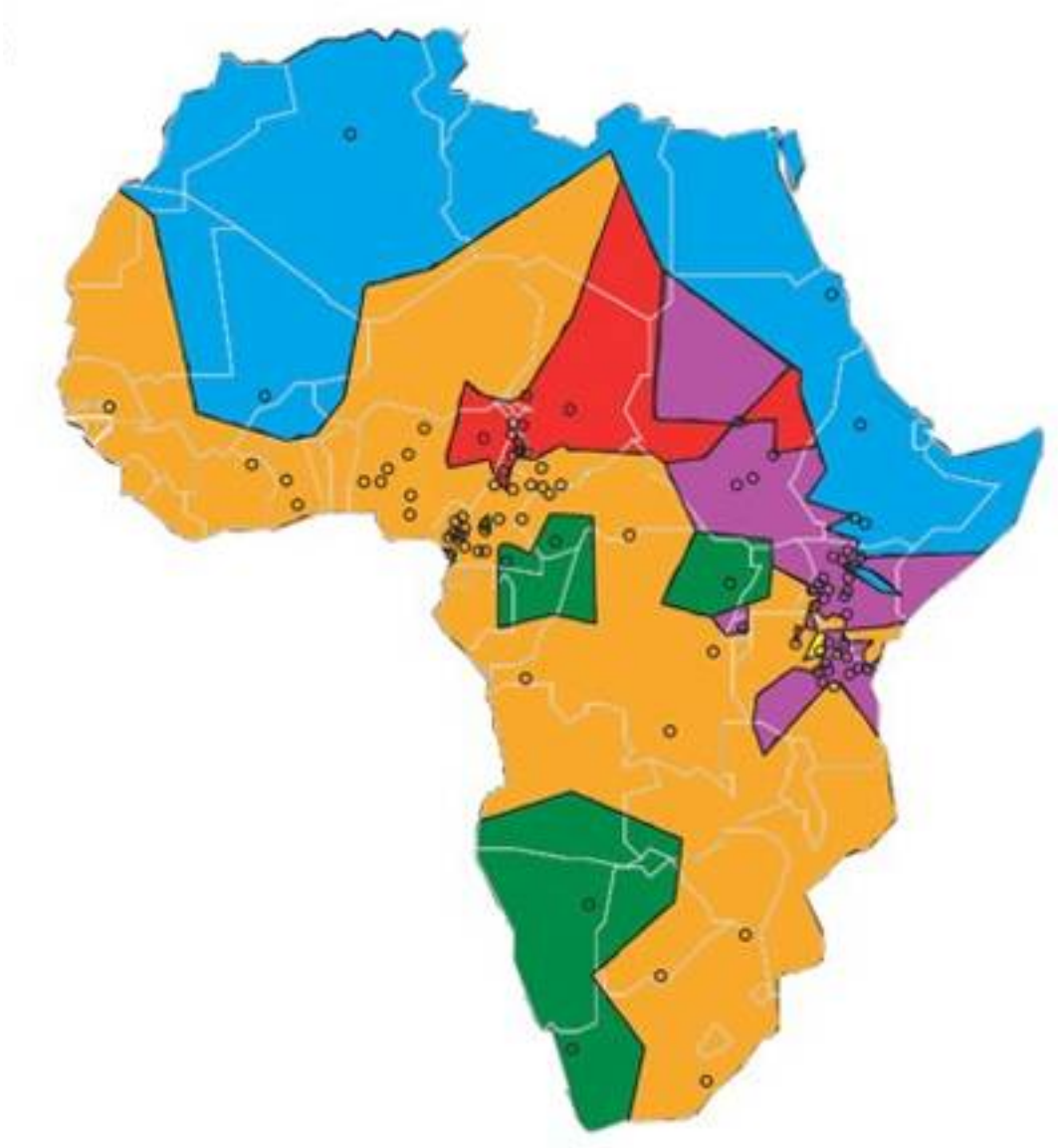


Figure 1.14 Map of Africa divided into six ethnic groups based on genetic structure
Genetic variation was assessed by analysis of 1327 microsatellite and indel markers in 121 African populations. Ethnic groups were determined using a Bayesian clustering algorithm assuming no admixture. Figure from [142].

Genetic population diversity decreases with increased distance from Africa [142], consistent with a series of founder-effect events occurring as populations formed from the out-of-Africa colonization waves. Figure 1.15 displays this graphically showing that genetic variation, as measured by a set of genome-wide polymorphic markers, is much reduced outside Sub-Saharan Africa.

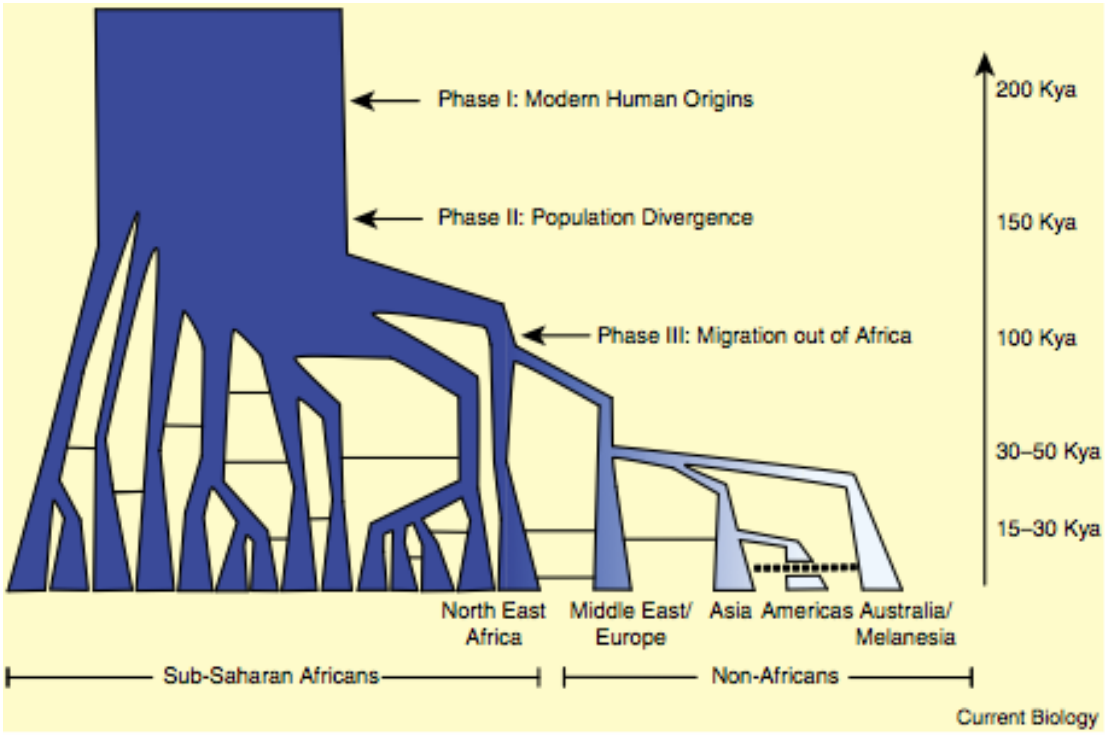


Figure 1.15 The evolution of worldwide genetic diversity over the last 200,000 years
This figure shows the worldwide population clusters, as determined from analysis of genome wide polymorphic markers, and the evolution of these clusters over time. A reduction is colour intensity indicates a reduction in genetic diversity. Horizontal black lines represent gene-flow between clusters. Figure from [143].

This pattern of distinct, separately evolving, African-offset populations is evident in the mitochondrial variation, with all modern humans traced back to an African founder. The position of this root within the mitochondrial phylogentic tree was originally confirmed with the aid of comparison to chimpanzee and Neanderthal sequences [52] and sits in the L super-haplogroup between the deep-routing branches L0 and L1-5 [144]: see Figure 1.16.

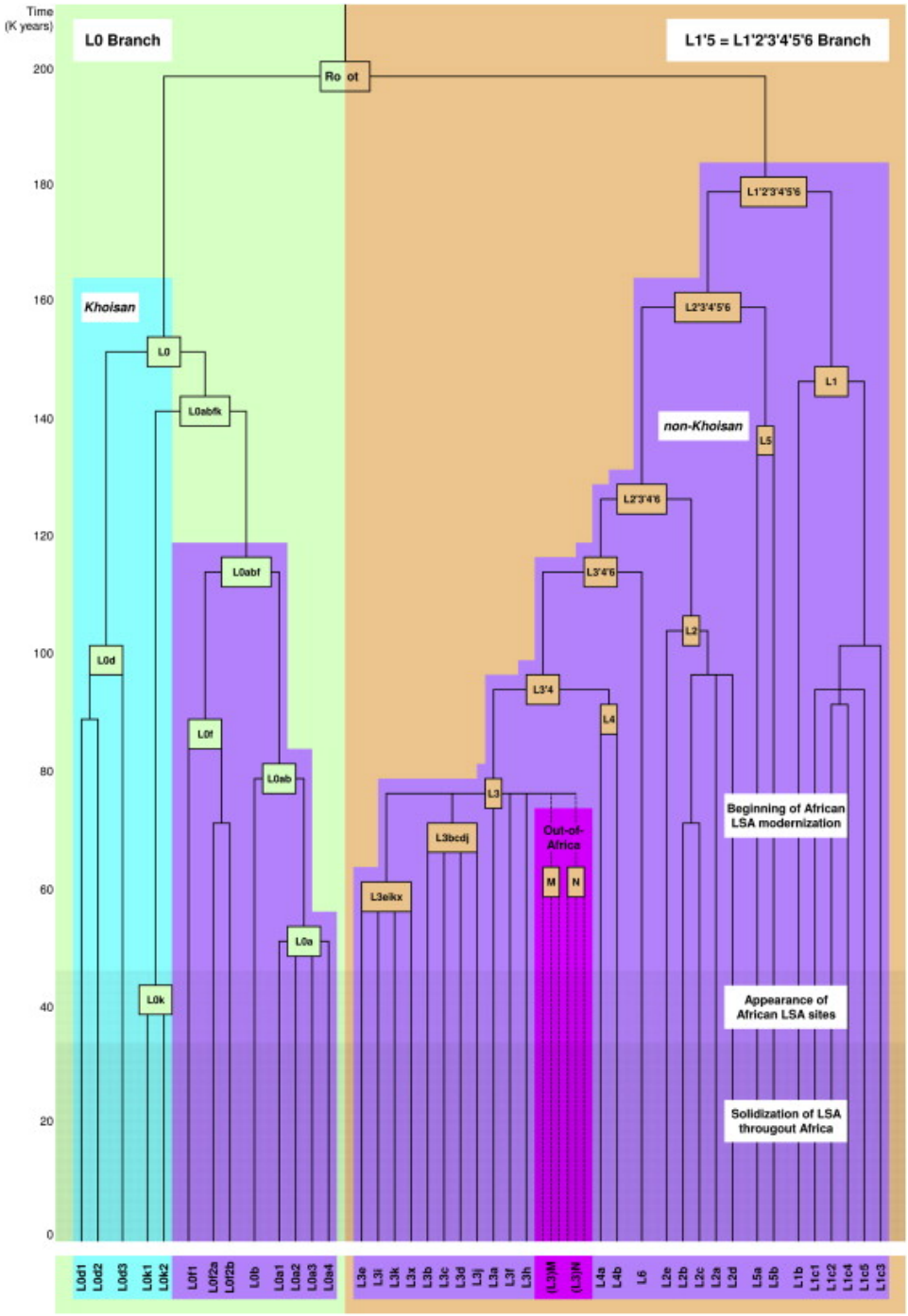


Figure 1.16 Known mitochondrial phylogenetic tree
Phylogenetic tree showing the evolutionary mitochondrial root (at the top) to which all present-day human mitochondrial genomes can be traced back to. The phylogenetic arrangement (and hence evolutionary relationship) between African haplogroups L0-L6 is shown, with the lineages giving rise to all non-African haplogroups shown in the centre in pink. Figure from [144]

Most African mitochondrial lineages belong to one of these six L haplogroup families [52]. The geographical distribution of these different L-haplogroups within sub-Saharan Africa is complex, partly due to the movement and assimilation of populations caused by various historical events and cultural developments. For

example L0a is common in East, Central and Southeastern Africa while being almost absent from North, West and southern Africa suggesting an Eastern origin for this haplogroup [52]. In comparison, the related haplogroups L0d and L0k are mainly restricted to South Africa, being highly enriched within the Khoisan people (as highlighted in blue in Figure 1.16) [144]. One significant influence on mitochondrial haplogroup distribution within sub-Saharan Africa was the Bantu expansion [52], driven at least partly by agricultural advances.

All indigenous non-African haplogroups descend from L3 giving rise to the macro-haplogroups N and M created by the major out-of-Africa colonization events (see Figure 1.17). M haplogroups are spread across South Asia (the sub-continent) [145], and haplogroups derived from M (C, D, E, G and Z) are found throughout East Asia along with some M branches (e.g. M7) [146]. N haplogroups can be seen across the non-African world, often with specific geographical distributions, and haplogroup N additionally gives rise to haplogroup S (observed mainly in Oceana), European enriched haplogroup X, the Amerindian and East Asian haplogroup A, as well as haplogroups I, Y and W [132, 147, 148]. Macro-haplogroup R is the major offshoot of macro-haplogroup N and gives rise to the majority of commonly observed European haplogroups, including U, J, T and H [132, 147]. Haplogroup H is by far the most common European haplogroup, being observed at a frequency of about 40-50% across the continent [149-151], with subtype haplogroup H1 accounting for 13% of the European gene pool [152]. The ubiquitous nature of this haplogroup across Europe also extends to neighbouring populations, with haplogroup H frequencies of about 20% in the Near East and Caucasus [151], over 25% in some North African populations [149], and even about 10% in some areas of Central Asia [152]. Mitochondrial DNA also provides other evidence of gene flow back into North Africa (above the climatic obstacle of the Sahara desert) with U6 found at an appreciable frequency in North Africa populations, most likely as a result of population movement back into Africa from the Near East (where U6 can still be found at low levels) about 50,000 years ago [153].

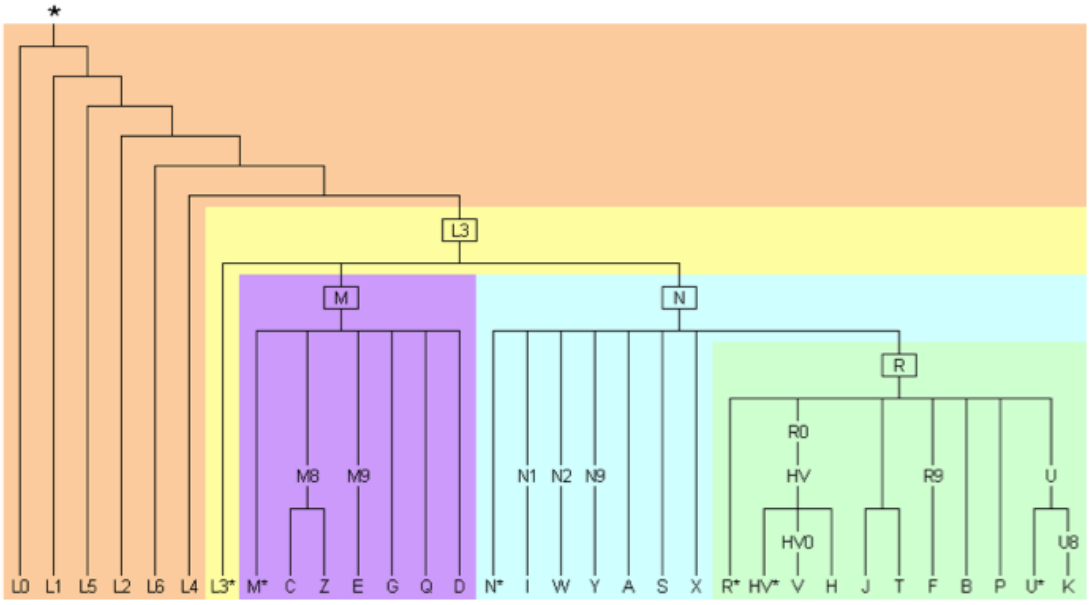


Figure 1.17 Simplified worldwide mitochondrial haplogroup phylogeny
All non-African haplogroups derive from the M and N macro-haplogroups that arise from the L3 branch. Macro-haplogroup N additionally gives rise to macro-haplogroup R from which most European mitochondrial genomes originate. Figure from [132].

Recent research has shown that the Neanderthal genome is more closely related to non-Africans than Africans, suggesting the presence of gene flow from Neanderthals to non-African humans during the time the two co-existed; accounting for 1-4% of present-day Eurasian DNA [154]. This genetic admixture is not detected in mitochondrial DNA, where Neanderthal sequences are significantly different from modern humans' [154].

1.4 Direct Phenotype Characterisation

An alternative approach to deducing geographic ancestry from an unknown DNA sample would be to define the individual's phenotype directly and bypass any inference gained from an ancestry prediction. In this way features such as eye colour, skin tone, hair colour and morphology, age, height, and facial structure could be directly determined from the DNA. This is an area of great interest at the moment, and progress has been made on some of these topics. It is now possible to predict eye colour with appreciable accuracy [155] and age within a 10-20 year window [115, 156, 157], while informative facial structure [158] and height prediction [159] are still well beyond the scope of current human genetic knowledge.

Despite these advancements, geographic ancestry information will still be a very useful investigative tool when building up a profile of an individual from an unknown DNA sample. A holistic phenotyping solution still remains a distant goal, and technical challenges remain regarding the implementation to forensic casework due to the large amount of genetic data that would need to be determined from compromised DNA.

1.5 Ethical Issues

DNA is private and unique to each individual, and as such, any analysis and subsequent inferences taken from this data must be carried out within an ethically sound framework. Testing of an individual's DNA without their consent is a criminal offense, with exceptions being granted for specific uses, for example when in a forensic context for human identification or criminal investigation. This does not give law enforcement organisations unrestricted license; indeed there is a consensus on specific restrictions about what is deemed appropriate which excludes medically important information [160] and instead focuses on externally visible characteristics [161].

Testing carried out within this project will be within a research framework using tissue from individuals who have consented to have their DNA analysed, with the findings of this research having multiple potential uses beyond forensics. Genetic characterisation of ethnically distinct groups within the UK has applications in both population genetics and stratification analysis for medical genetics, in addition to intelligence uses for forensic investigators. Further consideration of the ethical issues must be undertaken before implementing any test in a live case forensic scenario.

1.6 Aims

From the viewpoint of population genetics, it's interesting to characterise the UK population groups to understand the differing genetic influences underlying them and their similarities to other populations. With the explosion of interest in the use of

DNA testing to trace back individual ancestry, this is currently an area of research to which considerable resources are being dedicated worldwide. It is especially relevant to the haploid markers that allow ancestry to be traced back without recombination and provide a better understanding of population movements in history.

Additionally, it is impossible to assess the significance of tests using genetic markers if the allele/haplotype frequencies are not known, hence the construction of relevant databases is vitally important. Lastly there is great interest in using markers for population determination, and the final goal is to use the information collected about the three main ethnic groups within the UK to devise a highly accurate classification method that can determine the broad continental ancestry of an unknown DNA sample in forensically relevant situations within the United Kingdom.

It has been demonstrated above that differentiation between populations at the genetic level takes many different forms, and hence there are many candidate markers to choose from. If any final population determination test is to be used within a forensic laboratory, then it will be challenging to analyse CNVs with the available equipment, and impractical to analyse so many STRs on a limited amount of DNA. I will therefore focus this investigation on 3 techniques more suited to typically available forensic material: Y-STR analysis (using a limited number of STRs), mitochondrial DNA sequencing (utilizing the high copy number per cell), and a small well-chosen subset of autosomal SNPs (including some markers associated with the known pigmentation genes).

The aim of this project is therefore to genetically characterise the main ethnic groups of the British population, and use this information to produce a test of population determination that can be applied in forensic situations (i.e. a limited quantity and/or quality of DNA). Specific objectives are:

- To develop robust Y-STR assays comprising established and new Y microsatellite markers and genetically characterise these STRs
- To compile Y-STR haplotype databases for the three main UK populations along with an Irish Caucasian population for comparison
- To develop a Y-STR classification system for population determination

- To sequence, and determine the haplogroup of, mitochondrial DNA from the three main UK population groups
- For comparison, to sequence and haplogroup the mitochondrial DNA from associated Caribbean and Irish populations
- To assess the utility of using mitochondrial DNA for population determination
- To develop an autosomal SNP assay for population determination
- To validate this assay on samples from the three major UK population groups along with a British Chinese population
- To evaluate the three different population determination methodologies individually and in combination

2 Materials and Methods

2.1 Samples

All samples were originally collected for the purpose of paternity testing by the haematology laboratory at the Blizard Institute, Bart's & The London School of Medicine and Dentistry; a fully accredited paternity laboratory within the United Kingdom (UKAS ISO17025, UK Ministry of Justice). Samples were collected in line with the 'Department of Health Code of Practice and Guidance on Genetic Paternity Testing Services' and full consent was obtained for the testing performed. The surplus material remaining after the testing was complete was stored in the haematology tissue bank in compliance with accreditation requirements. During the course of this research, only samples collected prior to September 2006 were used. Authority for the use of these samples is given in Part 1, Section 9, and Part 3, Section 45, subsection 2 of the Human Tissue Act 2004 [162]. Ethics approval was obtained from Queen Mary, University of London Ethics Committee.

Individuals donated samples as either whole blood, or buccal cells in the form of a mouth swab. The ethnicity/ancestry information for each individual was self declared at the time of sampling.

2.2 DNA Extraction

Before DNA can be analysed it must first be extracted from cellular material; in this case that material being either blood or buccal cells. Two methods of DNA extraction were used throughout this research: the manually performed Chelex method yielding an average 200 ng of DNA at a concentration of approximately 1 ng/µl, or the Qiagen extraction procedure using the EZ1 robot (Qiagen) generating more concentrated DNA of 5-10 ng/µl in a similar 200 µl final volume.

2.2.1 Chelex

Chelex is a chelating resin in the form of small beads that remove from solution polyvalent metal ions such as magnesium. This results in the inactivation of damaging nucleases that would otherwise degrade the DNA. The method used is modified from the original protocol [163].

Cell lysis was achieved by incubating the cells in 1 ml of DNA free water for 20 min. The cells were added to the water in the form of either 4 µl of whole blood or by removing 3 mm from the end of a mouth swab. Following the incubation, the tube containing the water and lysed cells was spun at 14,000 g for 5 min, after which all the liquid was removed apart from the bottom 20 µl. Next, 180 µl of 5% w/v Chelex 100 (Sigma) suspension was added to the tube, which was then incubated at 56ºC for 20 min. A further incubation at 100ºC for 8 min then followed after which the tubes were again spun at 14,000 g for 5 min. The DNA was now ready for use and stored at 4ºC.

2.2.2 Qiagen

The BioRobot EZ1 DNA extraction system (Qiagen) provides an automated method of DNA extraction and purification utilizing magnetic bead technology. Cells are lysed and, in the presence of a chaotropic salt, DNA is subsequently bound to the silica surface of the magnetic particles. Particles are removed with a magnet and DNA is eluted after undergoing a series of washing steps.

The procedure for blood and buccal samples is slightly different. For blood samples, the EZ1 DNA blood kit (Qiagen) is used with the EZ1 DNA blood card (Qiagen). The card is inserted into the EZ1 robot and, following the on-screen instructions, reagent cartridges, tips with tip holders and elution tubes (all supplied with the kit) are inserted into the robot. Up to 6 samples can be extracted concurrently and 200 µl of whole blood for each sample is transferred to a supplied 2 mL sample tube that is also placed within the robot in the specified position. Purified DNA is produced approximately 20 minutes after the robot has been started.

Mouth swab samples are treated in a similar way, except that the EZ1 DNA tissue kit (Qiagen) and EZ1 DNA tissue card (Qiagen) are used instead, and there is an initial incubation step not present in the blood extraction method. This initial step consists of placing the end 3 mm of the swab into a sample tube containing 190 µl of lysis buffer and 10 µl of Proteinase K (both supplied with the EZ1 DNA tissue kit). The sample is then incubated at 56ºC for 20 min on a Thermomixer Comfort heated shaker (Eppendorf) with shaking at 1000 rpm. Following the incubation, the sample tube is placed in the robot following the on-screen instructions and extraction proceeds as for the blood samples. Further details of the extraction procedure can be found in the EZ1 DNA Handbook [164]. DNA is stored at -20ºC until needed.

2.3 **PCR Amplification**

The Polymerase Chain Reaction (PCR) is a revolutionary technique pioneered in the early 1980s. By amplifying small sections of DNA in an exponential reaction, PCR facilitates analysis of selected areas of the genome from minimal quantities of starting material. Small synthetic DNA oligonucleotides, called primers, bind to regions of the DNA surrounding the area of interest and provide a base from which DNA nucleotides can be added. In this way, using the sample as a template to instruct which base to add, extension from the primer proceeds, creating a copy of the desired region of DNA. Since DNA is a double stranded molecule, two primers (forward and reverse) are needed to copy any segment of DNA, one to copy each strand.

2.3.1 Primer Design

Primer design is an important component in the success of any PCR reaction. Ideal primer design criteria include similar annealing temperatures for all primers, typically around 60ºC [165], a 40-60% GC content and a length of about 20 bp. The secondary structure of the primer is also important, and hence the design considerations include avoiding long runs of guanine bases (it causes the primer to kink) or self-complimentarity within the primer causing hairpin structures to form. It is also important to check for possible annealing between different primers in the same reaction and with other areas of the target DNA as this can cause the production of

non-specific products and reduce the yield of the target PCR amplicon. Primer design throughout this research was aided by the software package Primer Express (Applied Biosystems) that could suggest various primer combinations based on the inputted sequence and the design criteria (primer length, amplicon length, GC content, annealing temperature, hairpin formation). To check for interaction between primers, an additional software program called AutoDimer [166] was used, and the NCBI BLAST algorithm (www.ncbi.nlm.nih.gov/blast) was used to check for primer complementarity to other areas of the genome.

Primers were ordered from biomers.net and were additionally HPLC purified if modified with a fluorescent dye.

2.3.2 PCR Optimisation

In any PCR reaction, there are various vital components:

Deoxyribonucleotide triphosphates (dNTPs) – The nucleotides (building blocks) needed to construct the many copies of the relevant section of DNA.
DNA Polymerase – The enzyme that catalyses the incorporation of dNTPs.
Sample DNA – The template to be copied.
Primers – Single stranded oligonucleotides identifying the region of the sample DNA to be copied.
Reaction Buffer – To control the environment in which the PCR components react, specifically the pH and ionic strength.
Magnesium Chloride – A divalent cation that encourages DNA annealing and is necessary for DNA polymerase activity.

PCR optimisation is a process by which PCR efficiency is increased by ensuring that all reaction conditions are favourable. This can be achieved by changing the thermal cycling conditions. Alternatively, primer, DNA, magnesium chloride, dNTP and DNA polymerase concentrations can be modified to produce higher or purer yields of the desired product.

Throughout this investigation, the standard PCR volume was 10 µl, and amplification was either performed in 0.2 ml thin walled microtubes (Anachem) or 96-well plates (Starlabs).

2.3.3 Thermal Cycling

The DNA polymerase normally used in PCR is *Taq*, from the bacterium *Thermus aquaticus.* This bacteria inhabits hot springs, and the enzyme will therefore be thermostable at the high temperatures needed to achieve DNA denaturation *in vivo*. The specific *Taq* polymerase used throughout the majority of this work is AmpliTaq Gold DNA Polymerase (Applied Biosystems). This version of the enzyme is specially modified to make it inactive at room temperature, and must undergo a 10-minute activation step at 95ºC before PCR cycling can commence.

DNA amplification occurs in 3 stages: denaturation, annealing and extension. The double stranded DNA has to be broken apart (denatured) in order for the primers to bind (anneal) and then sufficient time has to be allowed for the extension of the primers that flank the area of interest. These 3 stages – denaturation, annealing, and extension, all require different temperatures. The denaturation requires a high temperature of around 94°C, the annealing is primer specific, but is often in the order of 55°C to 60°C, and the extension step is typically performed at 72°C. By repeating this cycle many times it is possible to achieve huge amplification from very small amounts of DNA due to the fact that each new copy can act as a template for replication in the next cycle. A GeneAmp 9700 (Applied Biosystems) thermal cycler was used throughout his research to enact these cycling conditions.

A final extension step can optionally be added to the PCR program if split peaks (2 peaks 1 base pair apart) are being observed rather than a pure product. This is a phenomenon by which the *Taq* polymerase has a tendency to add an additional base (normally an adenosine) to the end of the PCR product. If this addition is not completed for all products, then split peaks are observed. To encourage addition in all amplicons, the reverse primer can be designed to have a G as the starting base and a final extension step (of anything up to an hour) can be included at the end of the PCR program [167].

2.4 STR Detection

Throughout the PCR process, the primers are incorporated into each of the DNA copies. By labelling one primer in each set (forward or reverse) with a fluorescent dye, it is possible to detect the amplified DNA product.

Amplicon size determination is critical in STR analysis. PCR products can be separated using electrophoresis because DNA is negatively charged and can be drawn to a positive charge. By running these DNA products in a medium that separates according to size, smaller DNA product will be eluted earlier than larger ones. If fragments of known size (a size standard) are run concurrently, then the exact size of the DNA fragment can be determined.

During this project an ABI-Prism-310 genetic analyser (Applied Biosystems) or an ABI-Prism-3130xl genetic analyser (Applied Biosystems) was used. These machines have the ability to analyse 96 samples separately by running each sample through a polymer filled capillary (the 310 has 1 capillary and runs 1 sample at a time, the 3130xl is a more advanced model with 16 capillaries running 16 samples at a time). DNA fragments of up to 500 base pairs will elute within approximately 30 minutes (depending on the polymer used). Towards the end of the capillary there is a small window, devoid of the protective sheathing, through which a laser is fired that excites the dyes attached to the primers. The emitted fluorescence is detected by a CCD (charge coupled device) camera and the results are saved to file.

To prepare samples for capillary electrophoresis, 1 μl of PCR product is mixed with 10 μl of ultra pure deionised formamide (Applied Biosystems) and 0.4 μl of ROX-GS500 internal size standard (Applied Biosystems) in a 310 MicroAmp reaction tube (Applied Biosystems) or 3130 96-well plate (Applied Biosystems). Septa (Applied Biosystems) are applied to cap the tubes or plates that are then heated for 5 min at 94°C to denature the DNA. The formamide prevents the DNA re-annealing.

The fluorescent dyes in use each emit fluorescence at different wavelengths, producing red, green, yellow and blue light. Although each dye has a wavelength of maximum fluorescence, they each additionally emit weaker fluorescence over a range of wavelengths. Due to this weaker fluorescence, at some wavelengths a detected signal could be due to more than one dye. Matrix files are used to compensate for this spectral overlap on the 310, while a spectral calibration must be performed on the 3130. During this project the dyes used were:

6-FAM	Blue
NED	Green
VIC	Yellow
ROX	Red

Standard injection and running conditions were used on both machines with respect to the polymer loaded. POP-4 polymer was used on the 310 while POP-6 (a slightly higher sieving polymer) was used on the 3130xl.

2.5 Sequencing

2.5.1 Sanger Sequencing

Sanger sequencing is a technique that determines the nucleotide sequence of a DNA region having first amplified that area with PCR. Irrespective of the specific PCR reaction performed, gel electrophoresis was carried out subsequent to PCR to check for successful PCR amplification before proceeding to the sequencing stage. PCR was carried out in a final volume of 20 µl.

2.5.1.1 Gel Electrophoresis

A 1.3% agarose gel was prepared by combining 1 g of analytical grade agarose (Promega) with 75 ml of 0.5x TBE (Sigma) and 7.5 µl of Gel Red stain (VWR). Eight microliters of each PCR product was loaded into a well on the gel (up to a maximum of 32 samples) and the gel was then run for 1 h at 100 V in 0.5x TBE

(Sigma). A size ladder (Hyperladder IV, Bioline) was also run on the gel. Once the electrophoresis was complete, the gel was viewed under UV light in an Alpha Imager gel dock system (Alpha Innotech Corporation) allowing the PCR product bands to be visualized and sized with respect to the size ladder. PCR is normally shown to have been successful by the presence of a single, strong band of the expected size.

In cases of multiple PCR products (specifically with reference to the Y-STR sequencing in chapter 3), the required fragment was first isolated on a 2.5% agarose gel, excised from the gel and the DNA recovered by centrifugation through glass wool. In cases of products of very similar size, separation was obtained by fractionation with a 6% polyacrylamide gel, staining with ethidium bromide and removing the appropriate portion. The DNA was eluted by crushing the gel portion and soaking overnight in water.

2.5.1.2 Sequencing Reaction Preparation

For samples demonstrating successful amplification, the remaining 12 µl of PCR product was combined with 2 µl of ExoSAP-IT (USB Corporation) and incubated in a thermal cycler at 37ºC for 1 h and 80ºC for 15 min. A negative control reaction was included with every batch of PCRs (by substituting water for sample) to ensure that there was no contamination at the PCR stage; in the event that the gel should show a band present from this negative control, all PCR reactions in that batch would be discarded.

Following enzymatic digestion of unincorporated dNTPs and primers with ExoSAP-IT, sequencing reactions were set up using the Big Dye Terminator v3.1 kit from Applied Biosystems. The sequencing reaction mix was diluted from the manufacturers protocol so that each 6.8 µl reaction contained 0.68 µl of Big Dye Terminator v3.1 Ready Reaction Mix (Applied Biosystems), 1.02 µl of Big Dye Dilution Buffer (Applied Biosystems), and a final concentration of 0.3 µM sequencing primer, with the remaining volume made up from water and purified PCR product (1 µl of PCR product was added as standard unless the band visualized on the gel was very weak, in which case more would be used).

The cycle sequencing program used to amplify these reactions consisted of an initial hold at 96ºC for 4 min, followed by 25 cycles of: 96ºC for 15 s, 50ºC for 10 s and 60ºC for 2 min. Excess unincorporated dye terminators were removed from the reaction products using a variation of the standard EDTA/ethanol procedure described in the Big Dye Terminator v3.1 guidelines [168]: due to the reduced reaction volume only 1.7 µl EDTA and 20 µl 100% ethanol were added to the sequencing products in the first stage of the clean-up. Sequencing products were separated by size on a 3130xl Genetic Analyser (Applied Biosystems) using POP-7 and a 36 cm capillary.

2.5.2 Pyrosequencing

For some samples it was necessary to carry out additional analysis of SNPs within either the mitochondrial DNA coding region or on autosomal chromosomes. Pyrosequencing technology was one method utilized to achieve this. Pyrosequencing is a real-time sequencing technique based on a luciferase reaction cascade.

A PCR was carried out initially to amplify the area of interest containing the SNP (or SNPs). PCR primers were designed using the PyroSequencing Assay Design Software 1.0.6 (Biotage), with amplicon lengths usually specified to be around 100 bp to ensure efficient amplification. One primer in every pair was labelled at the 5' end with biotin, and it was this strand incorporating the biotin that was eventually analysed by the pyrosequencer. PCR was carried out in 10 µl reactions containing 5 µl MyTaq HS Red Mix (Bioline), 1 µl DNA extract and a final concentration of 0.2 µM forward and reverse PCR primer. The assay design software attempts to produce primers with similar Tm's so that one universal set of thermal cycling parameters can be used. The technical notes suggest carrying out a 50 cycle PCR in order to exhaust all the biotinylated primer in the reaction, however there are too many issues with contamination at 50 cycles given the low DNA yields frequently encountered in forensic work, and so instead the reactions were run at 39 cycles. The PCR program used included an initial enzyme activation step of 95ºC for 5 min followed by 39 cycles of 95ºC for 15 s, 58ºC for 30 s and 72ºC for 30 s, with a final extension step of 5 minutes at 72ºC.

Following PCR, the success of the reaction could be checked visually on an agarose gel in an identical manner to that laid out in section 2.5.1.1. Alternatively, the amplicons were taken straight through to the pyrosequencing stage where a sequencing primer binds to the biotinylated strand allowing sequencing to commence. The primer usually binds 1-2 bp 5' of the SNP and pyrosequencing proceeds for a few bases through the region containing the SNP. The sequencing primer is designed by the assay design software at the same time as the PCR primers, and ideally has a Tm of 45-55ºC.

To enable the pyrosequencing primer to bind, the PCR product first had to be made single stranded. This is achieved by utilizing the biotin label to bind the PCR products to streptavidin beads: 10 µl of PCR product was combined with 2 µl Streptavidin Sepharose High Performance Beads (GE Healthcare), 38 µl PyroMark Binding Buffer (Qiagen) and 30 µl water and vortexed at 800 rpm for a minimum of 15 min. This was all carried out in a clear non-skirted 96-well plate (Starlab).

The beads were then captured using a vacuum prep workstation (Qiagen). This is a device with 96 prongs that is lowered into the 96-well plate and draws up the liquid by suction (Qiagen). The suction keeps the beads attached to the end of the filter probes and the non-biotinylated amplicon strand is stripped off by drawing 70% ethanol through the probes for 5 s, 0.1 M sodium hydroxide for 5 s and PyroMark Wash Buffer (Qiagen) for 10 s. Once the three washes have been completed, the device is positioned over a 96 well pyrosequencing plate (Qiagen), the vacuum is turned off, and the device agitated to ensure that all beads have been released into the appropriate wells. Contained within each well was 11.5 µl of PyroMark Annealing Buffer (Qiagen) and 0.5 µl of the relevant 10 µM sequencing primer. The plate was then heated to 80ºC on a heated block for 3 min before being placed in the Pyrosequencer (PyroMark MD, Qiagen).

The principal underlying pyrosequencing is that light is released when a nucleotide is incorporated into the newly forming sequence, and the amount of light released is proportional to the quantity of nucleotide incorporated (e.g. if the next bases in the sequence are two T's then this will produce twice as much light as one T would have). Each nucleotide is added to the reaction individually in a predetermined order, and if

the nucleotide added corresponds to the next nucleotide in the sequence then a DNA polymerase enzyme will incorporate this nucleotide into the sequencing product releasing pyrophosphate. The pyrophosphate reacts with the enzyme ATP sulfurylase in the presence of adenosine 5' phosphosulphate to produce ATP, and this ATP then interacts with luciferase to convert luciferin to oxyluciferin producing light. All four nucleotides, along with the enzymes and substrates (PyroMark GoldQ96 reagents, Qiagen), are added to the machine prior to the run, and are then automatically added to the wells of the plate by the instrument. The enzymes and substrates are added to the plate at the start of the run, and the software determines the order in which the nucleotides are to be added with reference to the known sequence (i.e. the flanking region of the SNP 3' to the sequencing primer), and any known bases where variation is expected (SNPs). Occasionally (especially just before and after the SNP of interest) a nucleotide will be added that is not expected to be incorporated (e.g. an A is added when the reference sequence dictates that the next base is expected to be a G); this is to ensure that there are no background interactions, and that the expected sequence is indeed being analysed.

The Pyrosequencing analysis software (Biotage) is used both to run the machine and analyse the results. Before a run can proceed, the details for each well are specified in the software – sample name and SNP assay. The SNP assay includes details about the sequence to analyse – the bases around the SNP of interest, the expected SNP alleles and the location of the sequencing primer, from here the software can determine in what order to add the nucleotides for each well in order to be most efficient. During a run a pyrogram is produced showing the intensity of any light produced following the addition of a nucleotide, and from this the SNP genotype can be determined.

Four controls are run for every new SNP assay to check for any unwanted interaction between the components. These consist of wells in the pyrosequencing plate containing annealing buffer and (a) the biotinylated primer only, (b) the biotinylated primer and the sequencing primer, (c) template only with no sequencing primer, and (d) template from a PCR negative with sequencing primer. If any background signal was seen on a pyrogram produced from one of these control reactions then the assay was redesigned.

2.5.3 Minisequencing – SNaPshot

Minisequencing is a technique that sequences just a single base and is used to determine the genotype of a SNP. Minisequencing can be implemented in various different forms, but SNaPshot (Applied Biosystems) has been used in this project. Interpretation of SNaPshot results can be complex, partly due to the unequal fluorescence intensity for A/G heterozygotes, however it does have the advantage of being able to analyse multiple SNPs in one reaction.

Minisequencing also requires an initial PCR to be performed. During development and optimisation of a new reaction these PCR products were visualised on an agarose gel as detailed in section 2.5.1.1 to check for clean and efficient amplification, however this step was not necessary for routine use. Excess dNTPs and unincorporated primers were removed from the PCR products by combining 1.3 µl ExoSAPit (USB Corporation) with 2.5 µl of PCR product and incubating at 37ºC for 1 h followed by 15 min at 80ºC.

SNaPshot reactions were carried out in 3 µl volumes containing 1 µl SNaPshot Ready Reaction Mix (Life Technologies), 1 µl ExoSAPit treated PCR product and 1 µl SNaPshot primers. Cycle sequencing parameters were 31 cycles of 96ºC for 10 s, 55ºC for 5 s and 60ºC for 30 s. The SNaPshot mix only contained dye labelled ddNTPs rather than the conventional dNTPs, and proceeds on the principal that an added SNaPshot primer will bind to a PCR product one base pair 5' to the SNP, and the minisequencing reaction will proceed by adding just one single base to this SNaPshot primer, that base changing depending on the SNP allele. Each ddNTP is labelled with a different colour dye, and so the base added (and hence the allele) determines the fluorescence emitted when the SNaPshot product is subjected to laser light.

Following completion of the SNaPshot reaction, any remaining ddNTPs were removed by adding 1.3 µl SAP (USB Corporation) to the SNaPshot reaction and incubating again at 37ºC for 1 h and 80ºC for 15 min. Products were separated by size and fluorescence on a 3130xl with a 36cm capillary (Applied Biosystems) running POP-7 (Applied Biosystems) on dye set E5. This was achieved by

combining 1 µl of SNaPshot product with 0.4 µl LIZ120 size standard (Applied Biosystems) and 10 µl Hi-Di formamide (Applied Biosystems).

Numerous SNPs can be analysed together following multiplex PCR by combining multiple SNaPshot primers in the same reaction. A typical SNaPshot primer would be 20 bp, but by adding a non-specific tail to the 5' end this product size can be increased so that multiple SNPs can be simultaneously visualised, e.g. SNP1 might have a product size of 20 bp while SNP2 could be 24 bp, SNP3 28 bp etc.

2.6 Analysis

2.6.1 Fragment Analysis

The resulting electropherograms from fragment analysis applications (either STR or SNaPshot analysis) were displayed and interpreted using either GeneScan v3.1 (Applied Biosystems) or GeneMapper ID v3.2 (Applied Biosystems) software. The software is able to provide precise sizes for sample peaks by comparison with the synthetic size standard fragments inserted into each well and hence integral to each profile.

For STR applications these sizes can then be converted into alleles (i.e. the precise number of repeats in an STR marker) by comparison with an allelic ladder. An allelic ladder is a collection of common alleles, each with a verified repeat number, and when run on the machine each known allele will migrate with a specific size, hence allowing comparison to the sizes obtained from the sample. An allelic ladder is run about once every 16 samples to correct for any subtle fluctuations in the instrument performance. Sample peaks were designated a specific allele number if matching to a peak in the allelic ladder by up to ±0.5bp.

For analysis of SNaPshot products, bins were setup in the software corresponding to the known elution size of the SNaPshot fragment for each SNP. The dye label effects the mobility of the product disproportionately for smaller sequence lengths and the size of the visualised fragment will vary depending on which nucleotide, and hence

fluorescent dye, is incorporated. For the same sized product, blue (G) peaks elute first followed by yellow (C) peaks, then green (A) peaks and finally red (T) peaks. SNP genotypes were confirmed by the presence or absence of the correct coloured peak within the expected size range (bin).

2.6.2 Sequencing Analysis

Sequencing analysis was carried out using Sequence Navigator (Applied Biosystems) for all Y chromosome work detailed in chapter 3. Sequences were first checked and edited for sequence coverage (i.e. the portion of sequence data to use) and miscalled bases. Forward and reverse sequences were then aligned manually and a consensus sequence generated. This consensus sequence could then be compared to a reference sequence or another sample.

For the mitochondrial sequencing analysis undertaken for chapter 4, sequencing products were analysed using SeqScape software v2.6 (Applied Biosystems). All sequences for any particular sample were imported into 1 project and an algorithm in the software compared these sequences to the known reference sequence – automatically reverse complementing sequences in the opposite direction and aligning all sequences with each other and to the reference. The Revised Cambridge Reference Sequence [80] was downloaded from GenBank (www.ncbi.nlm.nih.gov) where it is listed as the homo sapiens mitochondrial reference sequence (number NC_012920) and this was set in SeqScape as the reference genome. Any variations between this reference sequence and the consensus sequence from the sample were highlighted by the software and, after manual review of these bases, the list of these changes was exported.

2.7 Y Chromosome STR Investigation

Eleven Y Chromosome STR markers were used throughout this study: DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438 and DYS439. A widely used pentaplex reaction amplifying DYS19, DYS389I, DYS389II, DYS390 and DYS393 was used, and all reaction conditions

and primer sequences are as specified by Gusmão *et al.* [73]. During the research, the remaining markers were amplified under various multiplex combinations and reaction conditions. Primer sequences for these markers were as previously described (DYS437, DYS438 and DYS439 [74], DYS385 and DYS392 [70], and DYS391 [169]). A couple of additional markers, DYS460 and GATA A10 were also genotyped for some of the samples in the mutation study, and primer sequences for these loci are detailed in White *et al.* [75].

Full 11 marker Y-STR haplotypes were generated for 250 individuals of Caucasian, Afro-Caribbean and South Asian origin. All donors were resident in mainland Britain; Afro-Caribbean donors comprised of individuals sampled with ancestry in the Caribbean or Africa, while South Asian donors comprised of individuals with ancestry from Bangladesh, Pakistan or India. Additionally 155 unrelated White Caucasian individuals from the Republic of Ireland were genotyped and a set of 250 father:son pairs (with confirmed paternity indexes of over 1000) were typed to investigate the marker specific mutation rates (although full profiles were not achieved for all samples). In all cases where a Y-mutation was observed in a father:son pair, paternity was further confirmed by testing the trio (father, mother, child) with a 16 autosomal loci STR kit called PowerPlex 16 (Promega) according to the manufacturers protocol [170].

Sequencing of PCR products was necessary at various times throughout this research, both when producing allelic ladders and when investigating variant alleles and mutation events. The Y-SNP M48 was additionally analysed by direct sequencing in two samples using previously published primers [171].

Alternative PCR primers were designed for DYS390 and DYS385 a/b to further investigate non-standard alleles that were detected. For DYS390, these were: DYS390/1: TCA ATA TCA CAG AAC ATC GTA ATC CA, and DYS390/2: GAG ACA GTG TAT CCG CCA TGG TA. It was necessary to use a semi-nested PCR approach to analyse the DYS385 alleles individually. In the first stage, the DYS385a and DYS385b loci (see Figure 1.12) were amplified in separate reactions by using the standard DYS385 reverse primer with one of these new primers: DYS385a/F: TGT CAG AGA CTA GGA ATG CA or DYS385b/F: GTG GCT ATT GAG CAC TTG A.

These PCR products then underwent normal DYS385 amplification with standard primers in a second round of PCR. Primer concentrations of 0.15 µM were used for the first round and 0.3 µM for the second. PCR cycling conditions were as originally stated by Schneider *et al*. [172], with the modification that 9 fewer cycles were used in the first round. In all amplifications, standard PCR reaction conditions were employed: 200 µM each dNTP, 1.5 mM $MgCl_2$, 1x PCR Gold buffer (Applied Biosystems) and 0.5 U AmpliTaq Gold (Applied Biosystems) in a 10 µl reaction volume.

The specific identification of DYS385a and DYS385b alleles was only performed on a small selection of samples and hence DYS385 was excluded from the AMOVA analysis because for most individuals it was not possible to assess the precise differences between 2 samples (e.g. two individuals could share the same apparent 12,13 genotype but actually be DYS385 a/b 12,13 vs. 13,12). DYS389II alleles were represented without the DYS389I component (i.e. DYS389II-DYS389I) to accurately represent the changes occurring in only the DYS389II part of the marker.

2.7.1 Statistical Analysis

Gene and haplotype diversity values (a measure of how polymorphic individual markers or sets of marker are in a given population), along with the respective standard deviation figures, were calculated using the formulae described by Nei [173]. Allele and haplotype frequencies were estimated using the gene counting method (number observed/total sample number). These calculations were all accomplished in the Microsoft Excel files containing the collected population data. Confidence intervals of 95% were calculated for all derived mutation rates using the binomial distribution, and Fisher's Exact test was used to test for significant differences between these mutation rates.

More complex Analysis of MOlecular VAriance (AMOVA) calculations were performed using the Arlequin software [174], with genetic distance measured using R_{ST}. AMOVA provides a measure of how variance would change if population groups mixed freely. The results of this analysis can be summarized as R_{ST} values;

the larger the value, the greater the difference between the populations. A non-parametric permutational approach is implemented within the Arlequin software to assess whether this R_{ST} value represents a statistically significant difference between the two populations.

The Φst values and multi-dimensional scaling analysis relating to Y-STR European population divergence were carried out by Roewer *et al.* [175]. Phylogenetic networks were produced by applying a Median Joining algorithm [176] and subsequent MP calculation [177] to the haplotype data, as implemented in the software program Network 4.5 (www.fluxus-engineering.com). An example of the file input format is included in Appendix V.

A paternity index is a likelihood ratio stating how much more genetically likely it is that a specific man fathered a child compared with a random man selected from the same population. This is calculated by looking at the alleles shared between the father and child at a number of independent loci and the relative frequency of these alleles in the population. In this study paternity indices were calculated using previously established frequency databases within the laboratory and standard formulae [178].

2.8 Mitochondrial DNA Investigation

2.8.1 Control Region Sequencing

A mixture of male and female samples were sequenced for this study, some of the male samples being the same as those used in section 2.7 for the Y chromosome analysis. A total of 89 British Caucasian samples were tested, along with 135 Black British samples and 123 British Asian samples. Additionally 90 Irish Caucasian samples were typed as were 44 Jamaican samples, 44 Barbadian samples and a further 12 assorted Caribbean samples.

Initially, PCR reactions were carried out to enable analysis of the two mitochondrial hypervariable regions: HVI and HVII. These two areas of the mitochondrial genome

lie within the control region and show increased variability with respect to the rest of the genome. HVI spans the region 16,024 to 16400, while HVII lies further towards the end of the control region between bases 48-407. Standard primers were used to amplify these two areas – for HVI the L15977 [179] forward primer and H16401 [180] reverse primer were used, while for HVII the H408 [180] reverse primer was used with a L47 (CTC ACG GGA GCT CTC CAT G) forward primer. PCR was carried out in a 20 µl reaction with final concentrations of 200 µM each dNTP, 1.5 mM $MgCl_2$, 1x PCR Gold buffer (Applied Biosystems), 0.2 µM forward and reverse primers and adding 1U AmpliTaq Gold (Applied Biosystems) and 2 µl sample DNA. A 36 cycle PCR program was used (94ºC for 15 s, 56ºC for 1 min, and 72ºC for 2 min) with a 10 min initial *Taq* activation at 95ºC.

It is also possible to sequence the entire control region (approximately 1.1 kb), which would include HVI and HVII. This approach means that all of the bases in the control region are sequenced rather than just those within the hypervariable regions, and hence more information can be obtained about the mitochondrial sequence of the individual since sequence changes are regularly observed within these extra bases, albeit at a lower rate than in the hypervariable regions. Samples typed for just HVI and HVII were re-sequenced for the entire control region if subsequent data analysis showed that the information derived from the hypervariable regions was insufficient. Additionally, all samples tested towards the end of this study were sequenced for the full control region from the outset. The protocol for amplifying the entire control region was identical to that used for the hypervariable region PCRs, with the exception being that the primers used for the entire control region were L15977 with H599 [125].

It is expected that all samples should show a single sharp band corresponding to the correct sized PCR product (~450 bp for HVI, ~400 bp for HVII, and ~1.1 kb for the whole control region) when run on an agarose gel as detailed in section 2.5.1.1.

For HVI and HVII analysis, two sequencing reactions were prepared for each PCR product, one containing the forward PCR primer as the sequencing primer, and one containing the reverse PCR primer as the sequencing primer. Six sequencing reactions were performed for full control region analysis, these being the two PCR

primers (L15977 and H599) along with H16401, L16450, H274, and L314 [125]. Additionally, if heteroplasmy was found to be present due to changes in the poly-C stretch starting at 16184, then one additional sequencing reaction would be performed to provide an unambiguous sequence – H16175 [125].

Quality assurance for the mitochondrial sequencing regime was assured by successful completion of the EMPOP set of QC test samples (www.empop.org) and yearly ongoing quality control checks *via* the GEDNAP proficiency testing scheme [181].

2.8.2 Haplogroup Assignment

As more complete mitochondrial genomes are sequenced, so a more comprehensive understanding is being developed regarding the evolution of mitochondrial DNA throughout human history and migration. These mutations occur in a sequential way off a shared foundation, and were finally all collated into one mitochondrial phylogenetic tree in 2009 [132]. The Anderson reference sequence is located in the H2 branch of this phylogenetic tree, and it's possible to determine an individual's mitochondrial haplogroup by examining the differences from the Anderson sequence and working backwards along the tree from the H2 branch. In this way mitochondrial genomes from all individuals can be assigned into specific haplogroups, with all samples within this group having a shared ancestry. If assignment proved problematic (especially a case with some Caucasian samples where there can be relatively few differences to the Anderson sequence) then extra testing (either SNPs or whole genome sequencing) was carried out as detailed below in sections 2.8.3 and 2.8.4.

2.8.3 Additional SNP Analysis

Two suites of SNP reactions were developed to aid haplogroup assignment, the first tested a range of 8 SNPs that span the major worldwide haplogroups encountered in the UK, while the second set of 7 SNPs was specifically designed to help sub-classify haplogroup H individuals. In all cases, PCR was carried out in 10 µl reactions if analysing the SNPs directly on the pyrosequencer, or 20 µl reactions if first assessing PCR success by agarose gel electrophoresis. Details on the pyrosequencing technique

are given in section 2.5 while primer sequences are listed below in Table 2.1 along with the haplogroup significance of each marker. SNPs 10398 and 10400 were analysed in the same pyrosequencing reaction (they are only 2 bases apart), while SNPs 3915 and 3992 were analysed individually on the pyrosequencer from the same PCR product, and likewise with SNPs 4745, 4769 and 4793.

Table 2.1 Mitochondrial SNP primers for use with pyrosequencing

Oligo Name	Oligo Sequence	Haplogroup
mtDNA 3010 PCR F	TCACCAGTCAAAGCGAACTACTAT	
mtDNA 3010 PCR R*	AACCTTTAATAGCGGCTGCA	
mtDNA 3594 PCR F	TAGCTCTCACCATCGCTCTTCT	
mtDNA 3594 PCR R*	AAATAGGAGGCCTAGGTTGAGGT	
mtDNA 4216 PCR F	TCCTACCACTCACCCTAGCATTAC	
mtDNA 4216 PCR R*	TTGAGGGGGAATGCTGGA	
mtDNA 7028 PCR F*	AGCCCTAGGATTCATCTTTCTTTT	
mtDNA 7028 PCR R	AAGCCTCCTATGATGGCAAATACA	
mtDNA 10398/10400 PCR F	CTGGCCTATGAGTGACTACAAAAA	
mtDNA 10398/10400 PCR R*	GTCGAAATCATTCGTTTTGTTTAA	
mtDNA 12372 PCR F*	GCTTACGACCCCTTATTTACCGA	
mtDNA 12372 PCR R	GTGGTAAGGATGGGGGGAATTAG	
mtDNA 12705 PCR F*	GTTCGTTACATGGTCCATCATAGA	
mtDNA 12705 PCR R	GTTGGAATAGGTTGTTAGCGGTAA	
mtDNA 3010 PyroSeq	TGTTGGATCAGGACAT	H1
mtDNA 3594 PyroSeq	ATACCCAACCCCCTG	L (xL3/4)
mtDNA 4216 PyroSeq	TCACCCTAGCATTACTTATA	R2/J/T
mtDNA 7028 PyroSeq	TTGATAGGACATAGTGGAAG	H
mtDNA 10398/10400 PyroSeq	ACTACAAAAAGGATTAGACT	L/M, M/D/Z
mtDNA 12372 PyroSeq	GGGGGAATTAGGGAAG	U/K
mtDNA 12705 PyroSeq	ATTAGTATGGTAATTAGGAA	xR
mtDNA 3915/3992 PCR F*	CATCATGACCCTTGGCCATAA	
mtDNA 3915.3992 PCR R	TGTGTAGAGTTCAGGGGAGAGTGC	
mtDNA 4336 PCR F*	AGGAGCTTAAACCCCCTTATTTC	
mtDNA 4336 PCR R	TGGCACGGAGAATTTTGG	
mtDNA 4745-93 PCR F*	CTCTCCGGACAATGAACCATAAC	
mtDNA 4745-93 PCR R	TTGGGTAACCTCTGGGACTCA	
mtDNA 6776 PCR F	TCAATTGGCTTCCTAGGGTTTA	
mtDNA 6776 PCR R*	GAAATATGCTCGTGTGTCTACGTC	
mtDNA 3915 PyroSeq	GAGACTAGTTCGGACTC	H6a/H17
mtDNA 3992 PyroSeq	GGGTGTTTATTATAATAATG	H4
mtDNA 4336 PyroSeq	TGGGTTCGATTCTCAT	H5a
mtDNA 4745 PyroSeq	TTATGATTATTAATGATGAG	H13a1

mtDNA 4769 Pyro Seq	TTCCTAGTTTTATTGCTATA	H2
mtDNA 4793 PyroSeq	GAAGTGAAAGGGGGC	H7
mtDNA 6776 PyroSeq	TCGTGTGAGCACACC	H3

* indicates that the primer is biotin labelled at the 5' end. For each marker there are two PCR primers and a pyrosequencing primer. The haplogroup column shows the principal haplogroup(s) that this SNP is diagnostic for, but a change at this SNP could also be present deep within other branches - this is especially true of the rapidly mutating SNP at position 3010. Haplogroups prefixed with an 'x' denote that the SNP indicates the sample doesn't belong to this haplogroup, e.g. L(xL3/4) means that the sample belongs to haplogroup L, but does not fit within sub-branches L3 or L4.

2.8.4 Full Mitochondrial Sequencing

If the haplogroup assignment was still ambiguous following full control region sequencing and pyrosequencing SNP analysis, then in a few instances the entire mitochondrial genome was sequenced. PCR primer sequences were kindly supplied by Applied Biosystems and are listed below in Table 2.2; in total 46 different PCR reactions are undertaken for each full genome sequence. To simplify the sequencing step, each forward PCR primer is modified with a non-complementary 5' M13-F tail (TGTAAAACGACGGCCAGT) and each reverse PCR primer is modified with a non-complementary 5' M13-R tail (CAGGAAACAGCTATGACC). This results in each of the 46 PCR products having an additional short sequence at both ends of the amplicons that is identical between each product and independent of the sample DNA sequence. Hence, for the sequencing reactions, all PCR products can be sequenced using the same primers – an M13-F primer will give the forward sequence of each product and an M13-R primer will give the reverse sequence of each product.

The PCR, sequencing and associated processes were carried out in a similar way to that described previously in sections 2.8.1. Specifically the PCR conditions were modified to 40 cycles of 94ºC for 30 s, 60ºC for 45 s and 72ºC for 45 s with an initial 5 min incubation step at 96ºC and a final 10 min hold at 72ºC. Additionally the primer concentration of the forward and reverse primers in each of the 46 PCRs was slightly higher than that used in section 2.8.1 at a final value of 0.75 µM. As stated above, the sequencing reactions were carried out using M13 forward and reverse primers for each PCR product. Sequences were analysed with SeqScape v2.6 software and all 92 sequencing reactions were simultaneously aligned to the entire Revised Cambridge Reference Sequence; any differences were noted.

Table 2.2 List of PCR primers for full mitochondrial sequencing

Oligo Name	Oligo Sequence
RSA001145148 F	TGTAAAACGACGGCCAGTCCCGTTCCAGTGAGTTCACCC
RSA001145163 F	TGTAAAACGACGGCCAGTGGTTGGTCAATTTCGTGCCAG
RSA001145139 F	TGTAAAACGACGGCCAGTTGGCGGTGCTTCATATCCCTC
RSA001145117 F	TGTAAAACGACGGCCAGTGCCCGTCACCCTCCTCAAGT
RSA001145118 F	TGTAAAACGACGGCCAGTAACTTTGCAAGGAGAGCCAAAGC
RSA001145119 F	TGTAAAACGACGGCCAGTGCGTTCAAGCTCAACACCCA
RSA001145120 F	TGTAAAACGACGGCCAGTGCGGTACCCTAACCGTGCAA
RSA001145121 F	TGTAAAACGACGGCCAGTCCCTAGGGATAACAGCGCAATCCT
RSA001145122 F	TGTAAAACGACGGCCAGTCATACCCATGGCCAACCTCCT
RSA001145172 F	TGTAAAACGACGGCCAGTCCTCTAGCCTAGCCGTTTACTCAATCC
RSA001250242 F	TGTAAAACGACGGCCAGTCTTCGACCTTGCCGAAGGG
RSA001145173 F	TGTAAAACGACGGCCAGTCACCCCATCCTAAAGTAAGGTCAGC
RSA001250243 F	TGTAAAACGACGGCCAGTCTTCTGAGTCCCAGAGGTTACCCA
RSA001308205 F	TGTAAAACGACGGCCAGTTGGGCCATTATCGAAGAATTCACA
RSA001145177 F	TGTAAAACGACGGCCAGTCAGCTAAGCACCCTAATCAACTGGC
RSA001145178 F	TGTAAAACGACGGCCAGTCAGCTCTAAGCCTCCTTATTCGAGC
RSA001250244 F	TGTAAAACGACGGCCAGTTGCCATAACCCAATACCAAACGC
RSA001250245 F	TGTAAAACGACGGCCAGTCAATTGGCTTCCTAGGGTTTATCGTG
RSA001307967 F	TGTAAAACGACGGCCAGTCCCGATGCATACACCACATGAA
RSA001250246 F	TGTAAAACGACGGCCAGTGAGCTTATCACCTTTCATGATCACGC
RSA001145184 F	TGTAAAACGACGGCCAGTCTACGGTCAATGCTCTGAAATCTGTG
RSA001145185 F	TGTAAAACGACGGCCAGTGAAAATCTGTTCGCTTCATTCATTGCC
RSA001145124 F	TGTAAAACGACGGCCAGTCCTCCTCGGACTCCTGCCTC
RSA001145137 F	TGTAAAACGACGGCCAGTATTGGAAGCGCCACCCTAGC
RSA001145188 F	TGTAAAACGACGGCCAGTCGATACGGGATAATCCTATTTATTACCTCAG
RSA001145189 F	TGTAAAACGACGGCCAGTCGAGTCTCCCTTCACCATTTCCG
RSA001250247 F	TGTAAAACGACGGCCAGTCATTTTGACTACCACAACTCAACGGCTAC
RSA001145125 F	TGTAAAACGACGGCCAGTCTAGTCTTTGCCGCCTGCGA
RSA001145126 F	TGTAAAACGACGGCCAGTCCAACGCCACTTATCCAGCG
RSA001145192 F	TGTAAAACGACGGCCAGTCTTATGACTCCCTAAAGCCCATGTCG
RSA001145193 F	TGTAAAACGACGGCCAGTCAAACTACGAACGCACTCACAGTCG
RSA001145138 F	TGTAAAACGACGGCCAGTGGGCTCACTCACCCACCACAT
RSA001145127 F	TGTAAAACGACGGCCAGTTTACCACCCTCGTTAACCCTAACAAA
RSA001145196 F	TGTAAAACGACGGCCAGTCCTTCTTGCTCATCAGTTGATGATACG
RSA001145128 F	TGTAAAACGACGGCCAGTGCAGCAGTCTGCGCCCTTAC
RSA001250248 F	TGTAAAACGACGGCCAGTCCACATCATCGAAACCGCAAAC
RSA001145129 F	TGTAAAACGACGGCCAGTCAGCCCTCGCTGTCACTTTCC
RSA001145144 F	TGTAAAACGACGGCCAGTACGCCCATAATCATACAAAGCCC
RSA001145130 F	TGTAAAACGACGGCCAGTGCCATCGCTGTAGTATATCCAAAGACA
RSA001145131 F	TGTAAAACGACGGCCAGTCGCCTGCCTGATCCTCCAA
RSA001145132 F	TGTAAAACGACGGCCAGTGACAGTCCCACCCTCACACGA
RSA001145133 F	TGTAAAACGACGGCCAGTCTAGGAGGCGTCCTTGCCCT
RSA001250252 F	TGTAAAACGACGGCCAGTGAAAAAGTCTTTAACTCCACCATTAGCACC
RSA001250250 F	TGTAAAACGACGGCCAGTCCCCCCATGCTTACAAGCAAGT

RSA001145161 F	TGTAAAACGACGGCCAGTCAGGTCTATCACCCTATTAACCACTCACG
RSA001250241 F	TGTAAAACGACGGCCAGTTGGCCACAGCACTTAAACACATCTC
RSA001145148 R	CAGGAAACAGCTATGACCCCCAGTTTGGGTCTTAGCTATTGTGTG
RSA001145163 R	CAGGAAACAGCTATGACCCTGCTAAATCCACCTTCGACCCTTAAG
RSA001145139 R	CAGGAAACAGCTATGACCCGCCAGGTTTCAATTTCTATCGC
RSA001145117 R	CAGGAAACAGCTATGACCGGGATTTAGAGGGTTCTGTGGGC
RSA001145118 R	CAGGAAACAGCTATGACCGCATGCCTGTGTTGGGTTGA
RSA001145119 R	CAGGAAACAGCTATGACCGCAGGTTTGGTAGTTTAGGACCTGTG
RSA001145120 R	CAGGAAACAGCTATGACCGGGAAGGCGCTTTGTGAAGTAGG
RSA001145121 R	CAGGAAACAGCTATGACCGCGGTGATGTAGAGGGTGATGG
RSA001145122 R	CAGGAAACAGCTATGACCCGGTTGGTCTCTGCTAGTGTGGA
RSA001145172 R	CAGGAAACAGCTATGACCGTGTATGAGTTGGTCGTAGCGGAATC
RSA001250242 R	CAGGAAACAGCTATGACCGCGCTGTGATGAGTGTGCCT
RSA001145173 R	CAGGAAACAGCTATGACCGTTTGGTTTAATCCACCTCAACTGCC
RSA001250243 R	CAGGAAACAGCTATGACCAGGTAGGAGTAGCGTGGTAAGGGC
RSA001308205 R	CAGGAAACAGCTATGACCAGAGAATAGTCAACGGTCGGCG
RSA001145177 R	CAGGAAACAGCTATGACCGGCCTCCACTATAGCAGATGCG
RSA001145178 R	CAGGAAACAGCTATGACCCTGTTAGTAGTATAGTGATGCCAGCAGCTAGG
RSA001250244 R	CAGGAAACAGCTATGACCCTTCCGTGGAGTGTGGCGAG
RSA001250245 R	CAGGAAACAGCTATGACCGGGCATCCATATAGTCACTCCAGG
RSA001307967 R	CAGGAAACAGCTATGACCCTAGGATGATGGCGGGCAGG
RSA001250246 R	CAGGAAACAGCTATGACCGCTAAGTTAGCTTTACAGTGGGCTCTAG
RSA001145184 R	CAGGAAACAGCTATGACCGTCATTGTTGGGTGGTGATTAGTCG
RSA001145185 R	CAGGAAACAGCTATGACCGGTGGCGCTTCCAATTAGGTG
RSA001145124 R	CAGGAAACAGCTATGACCTGAGGAGCGTTATGGAGTGGAAG
RSA001145137 R	CAGGAAACAGCTATGACCCAGGTGATTGATACTCCTGATGCGA
RSA001145188 R	CAGGAAACAGCTATGACCTTATACTAAAAGAGTAAGACCCTCATCAATAGATGG
RSA001145189 R	CAGGAAACAGCTATGACCGGGTAAAAGGAGGGCAATTTCTAGATC
RSA001250247 R	CAGGAAACAGCTATGACCAGGCCATATGTGTTGGAGATTGAGA
RSA001145125 R	CAGGAAACAGCTATGACCGGGAAGGGAGCCTACTAGGGTGT
RSA001145126 R	CAGGAAACAGCTATGACCTGTCGTAGGCAGATGGAGCTTG
RSA001145192 R	CAGGAAACAGCTATGACCGTGATATTTGATCAGGAGAACGTGGTTAC
RSA001145193 R	CAGGAAACAGCTATGACCGTCGTAAGCCTCTGTTGTCAGATTCAC
RSA001145138 R	CAGGAAACAGCTATGACCTGGGTTGTTTGGGTTGTGGCT
RSA001145127 R	CAGGAAACAGCTATGACCCTGCTAGGAGGAGGCCTAGTAGTGG
RSA001145196 R	CAGGAAACAGCTATGACCGCTTTGAAGAAGGCGTGGGTACAG
RSA001145128 R	CAGGAAACAGCTATGACCGCTGCCAGGCGTTTAATGGG
RSA001250248 R	CAGGAAACAGCTATGACCGATGAGTGGGAAGAAGAAAGAGAGGAAG
RSA001145129 R	CAGGAAACAGCTATGACCGGATTGGTGCTGTGGGTGAAA
RSA001145144 R	CAGGAAACAGCTATGACCGGGAGGTCGATGAATGAGTGGT
RSA001145130 R	CAGGAAACAGCTATGACCAGGCCTCGCCCGATGTGTAG
RSA001145131 R	CAGGAAACAGCTATGACCGAAGGAAGAGAAGTAAGCCGAGGG
RSA001145132 R	CAGGAAACAGCTATGACCCGGATGCTACTTGTCCAATGATGG
RSA001145133 R	CAGGAAACAGCTATGACCGGGTTTGATGTGGGTTGGGTT
RSA001250252 R	CAGGAAACAGCTATGACCGGGAACGTGTGGGCTATTTAGGCT
RSA001250250 R	CAGGAAACAGCTATGACCCTGTGTGGAAAGCGGCTGTG

RSA001145161 R	CAGGAAACAGCTATGACCGGGTTGTATTGATGAGATTAGTAGTATGGGAG
RSA001250241 R	CAGGAAACAGCTATGACCCTATTGACTTGGGTTAATCGTGTGACC

2.8.5 Network Analysis

Sequence changes were collated on a Microsoft excel spreadsheet such that one column represented each base at which there was a change anywhere within the 537 samples and each row represented one sample, i.e. if just looking at HVI then there could be a different column for every base from 16024-16400, however bases where every sample contained the same nucleotide were excluded as they provide no useful information. The relevant nucleotide was recorded in this grid for each sample – for any one sample the vast majority of bases will match that listed in the reference sequence. Using the concatenate function in excel this grid was then turned into one long string of characters, i.e. ACTGTC etc., for each sample. The character strings for each sample were then copied into a new spreadsheet in column B from row 3 downwards. In cell A1 of this new spreadsheet was a list of every base that corresponded to a nucleotide in the string, with each base separated with a semicolon, i.e. if the first variant base in the dataset was 16037 then this would be recorded in cell A1 followed by a semicolon and then the next base, e.g. 16039. Cell A2 recorded the importance (weight) associated with each base (separated by semicolons), allowing changes at bases known to be hyper-mutable to be accorded less significance when developing the phylogenetic network; the default used was 10 for each base.

In column A from row 3 down, matching the sequence string in column B, was recorded some basic information about each sample. This was in the form:
>A;1;;;;B;C;D;;
where A is sample name, B is deduced haplogroup name, C is simple haplogroup (e.g. H rather than H1a2b3) and D is the population that the sample comes from. This spreadsheet was saved as a comma separated values (csv) file with a file extension of .rdf. On opening in Microsoft word, the find and replace function was used to replace all commas with a carriage return and the file re-saved. A simple example file is included in Appendix VI.

The file was imported into Network 4.6.1 (www.fluxus-engineering.com) and phylogenetic networks were computed by applying a Median Joining algorithm [176]. A maximum parsimony (MP) calculation [177] was applied to reduce complexity and the resulting phylogenetic tree visualised using Network Publisher software v1.3 (www.fluxus-engineering.com).

Through both experimentation with the quality of phylogenetic network produced and known mutation characteristics for certain nucleotides, bases 16182, 16183, 152 and insertions following positions 524 and 573 were all accorded a reduced weight of 5, bases 16311 and 195 were given a weight of 8, base 16519 and the deletion of bases 523 and 524 were given a weight of only 2 while 16223 was given a weight of 20 and any changes in the coding region (e.g. the SNPs typed by pyrosequencing) were given weights between 70 and 100. Insertions related to length heteroplasmy following the poly-C stretches at either 16193 or 315 were ignored as were point heteroplasmies.

2.9 Population Specific SNP Investigation

2.9.1 SNP Selection

SNPs were selected in the first instance bases on allele frequency differences between populations. This included polymorphisms showing, in the best-case scenario, completely opposing allele frequencies (i.e. a frequency of the C allele at 1 in population A and 0 in population B), those with one allele confined to a specific population, and those with skewed allele distribution (e.g. a C/T frequency ratio of 20:80 in population A but 80:20 in population B). A selection of tri-allelic SNPs were also chosen that showed differing distributions of the three alleles in different populations. SNPs were chosen on their ability to separate two of Caucasian, African and East Asian populations. Allele frequency data came from the published literature, the NCBI SNP database dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/) and the Applied Biosystems assay-on-demand SNP database (https://myscience.appliedbiosystems.com/common/search.jsp?assayType=genotyping).

Following selection of a set of candidate SNPs, SNP sequence data was downloaded from either dbSNP or the now obsolete Celera Genomics SNP Reference Database (previously located at http://www.celeradiscoverysystem.com). Sequence context was checked for the presence of clustering SNPs that may interfere with efficient amplification, and hence should be avoided during the primer design stage, *via* querying the aforementioned Celera database and the Santa Cruz USCS Genome Browser (http://genome.ucsc.edu).

The design of this initial population specific SNP panel was a collaborative effort across a number of different laboratories with SNP selection carried out by Chris Phillips in Santiago de Compostella, SNP sequence collation and clustering SNP identification carried out by me, multiplex PCR design carried out by Juan Sanchez in Copenhagen and assay validation and population testing performed by all three laboratories in addition to the forensic laboratory at Cologne University.

Over the course of a 6 month trial phase, genotype data on different populations from the four laboratories led to reassessment of the panel composition and exchange of some poorly performing SNPs with better performing ones. The initial panel was composed of only 22 SNPs but eventually reached 34. Additional Asian specific SNPs were selected, assays designed and genotypes generated exclusively at the London laboratory to try to reduce the Caucasian/South Asian misclassification rates.

Bauchet *et al.*[182] conducted a study looking at roughly 10,000 SNPs from nearly 300 individuals and produced a core list of those displaying divergent allele patterns between Northern and South-Eastern European samples. The English samples in the study clustered with the northern group, and if the south-eastern cline continues into Asia then it can be postulated that these SNPs might be useful for distinguishing European and South Asian individuals. Of those SNPs showing the highest diversity in this list, global allele frequencies obtained from dbSNP showed that most of these markers only presented negligible European-East Asian allele frequency differences (as with the vast majority of markers on dbSNP there are no South Asian specific allele frequencies). An exception to this was SNP rs942793 which showed the 2nd highest F_{ST} value in the study between the Northern and South-Eastern European populations while also demonstrating continental variation between those European

(0.675/0.325) and East Asian (0.294/0.706) genotypes submitted to dbSNP (frequencies refer to A/C alleles), hence indicating that this SNP might be an interesting candidate SNP to validate for any European-South Asian difference.

Seldin *et al.* [35] used an Illumina genome-wide SNP panel, consisting of over 5,700 markers, to test for population structure primarily in Europe, but they also tested a set of South Asian samples as a comparison to test for admixture. Ranking these SNPs bases on the South Asian-Western European delta value, the top 5 SNPs were chosen for testing here. Yang *et al.* [183] undertook a more focussed SNP selection methodology, and tested 195 SNPs on nearly 800 samples from across the globe. In a similar manner to above, SNPs were ranked by their South Asian-Western European delta values, and four of the most informative SNPs selected for further testing. Finally rs16891982 was selected for testing due to it's known involvement with hypo-pigmentation (light skin colour) in Europe [184, 185].

2.9.2 Genotyping

Due to the different iterations of the 34-plex SNP set, with different markers added and removed during the development and validation process along with primer mix adjustments to provide better balance to the final mix, the conditions given below are those finally settled on and are slightly different to those published [186]. The primer sequences for those SNPs that ended up in the final marker set are given below in Table 2.3 and Table 2.4 along with the final concentration of each primer in the PCR or SNaPshot reaction. PCR was carried out in a 6.9 µl reaction consisting of 1x AmpliTaq Gold Buffer (Applied Biosystems), a final concentration of 5.9 mM $MgCl_2$ (Applied Biosystems), 600 mM each dNTP (GE Healthcare) and 0.16 mg/ml non-acetylated BSA (Ambion), 0.5 U AmpliTaq Gold, 1 µl extracted DNA and concentrations of the 68 primers as indicated in Table 2.3. Thermal cycling consisted of an initial enzyme activation step of 95ºC for 15 min followed by 30 cycles of 95ºC (15 s), 60ºC (50 s) and 65ºC (40 s) before a final 15 min extension step at 75ºC. SNaPshot reactions were performed as detailed in section 2.5.3.

Table 2.3 PCR Primers and final primer concentration for the 34-plex SNP PCR

	SNP	PCR F	PCR R	Final concentration µM
1	rs2304925	CCCATTAACTCATCAAAGTGGTGAT	CCCCACTCCACCGCTAAT	0.201
2	rs5997008	GTCAACACTAGAGTATTTGCCCATC	ACAAACCCAAAGACTGTTCTGC	0.095
3	rs1321333	GTCAGTAAGACGGTAACTCC	CTAACACAAGCCTAAATCCAG	0.138
4	rs2814778	AACCTGATGGCCCTCATTAGT	ATGGCACCGTTTGGTTCAG	0.060
5	rs917118	GCCCTTTAGGGTCGGTTC	GTAAGAGATGACTGAGGTCAACGAG	0.050
6	rs1024116	CCATGTGTTCTAATAAAAAGGATTGC	TGGGAAGTGAGCAAAAGTAAATACA	0.150
7	rs7897550	CGATGTGTCTTACGGAATACTAGGT	AGAGCTGACAGGCAAAAATGCTAT	0.100
8	rs722098	GGAAGTACACATCTGTTGACAGTAATGA	GGGTAAAGAAATATTCAGCACATCC	0.115
9	rs10843344	TGTACAATGGTAGATGTGTGCTCAG	GATAGCTCTGGTGTTGCATTATTGT	0.150
10	rs239031	TAGCTGTGAGATAGAAATCCTGGAC	ACTACCCTAATCTCAGCTTCCACTC	0.025
11	rs12913832	ACGTTGGATGCGAGGCCAGTTTCATTTGAG	ACGTTGGATGAAAACAAAGAGAAGCCTCGG	0.150
12	rs1978806	AGAGTTTGACATGATGGTGCTCTA	TCTTGTTTCTAAGCAGGAAAGTTG	0.050
13	rs2040411	TCTGGAATGCCAGTTCTTTTGT	CAGAACGCCTATGAAAACCAGT	0.090
14	rs773658	ACAAACGGAAAGTAGTATTGGACTG	AGAAGGGGCACAGCAATTTAGTA	0.160
15	rs10141763	ACAGACTTGGTTCCCTGAAGTCTA	GTAGATTGTAGGCAAGTCGTAAAGG	0.150
16	rs182549	AAGTACTGGGACAAAGGTGTGAG	AGAAGTCAGAATACCCCTACCCTAT	0.311
17	rs1573020	CTATCTGCCACCTGAGAGAGTATTG	AGGTGTCAGCTTCTTCTGACCAT	0.025
18	rs896788	GTAATGCCTCTGTGGCCCTAT	ATTCCGTCCACATCTTCACTG	0.075
19	rs2065160	AAGAATGGCCTCTCGATGAGTA	GATGATACCTACGCATAGTCTGTTTACTTC	0.040
20	rs2572307	GTGTAGCTATGCCATCATTCAATC	ATCCTTAGAAGGGTGCTAAACTGAG	0.075
21	rs2303798	CCAGCCTGCACCACTGTC	AGAGATGTGTTCAGGAAGAGGCTA	0.216
22	rs2065982	CTTGGGGCAGTCTTTAAGTCCT	AGGAAGTGGTCAGTGCCAGTAG	0.100
23	rs3785181	CTCTGTTCAGTTTCAAAGTTCTGG	TTGTGTTCAAAAATTTCAATTAGGTT	0.063
24	rs881929	AGCTACCTGGTGTCTAACTC	TTGACCCAGTGGTTCTGAGC	0.150
25	rs1498444	GGCTATTACCACATTAAGAGAAACTGC	CAGCCTCTCAATGCAAATGAT	0.251
26	rs1426654	GAAGACTTTTTGCAAGCACGAG	GGCAAATGCTGTAAGAATCCAT	0.150
27	rs2026721	AATTCAGGAGCTGAACTGCC	TGTTCAGCCCTTGGATTGTC	0.160
28	rs4540055	TGTGCCTCTGATCACTTTTGAATAC	CCTAGCCAACTCCAGAGTTCAT	0.060
29	rs16891982	GAATAAAGTGAGGAAAACACGGAGT	GTTTCTCATCTACGAAAGAGGAGTC	0.060
30	rs1335873	GTGGATGATATGGTTTCTCAAGG	TTCAACAAACGTGTGATGCTCT	0.038
31	rs1886510	GTCCTTGTCAATCTTTCTACCAGAG	GGATTTTCACAACAACACTTGC	0.075
32	rs730570	CAGCACCCTGTAAAGTCCAG	CAGCACTCACCTGCATCTCA	0.065
33	rs5030240	CCAAAGTGCCAGGATCACAG	TCCCTAGAAATCCTTCAGCC	0.100
34	rs727811	GTGTTTCTTTTTCTCTTACCGGAAC	GTGAATGAAATCATGAGATTGCTG	0.100

Table 2.4 SNaPshot primers and final concentration for 34-plex assay

	SNP	SNaPshot Primer	Extension Size	Final concentration µM
1	rs2304925	CCACTCCACCGCTAAT	17	0.287
2	rs5997008	CAGGATCGATTGGTTCC	18	0.149
3	rs1321333	AAGACGGTAACTCCATGGCTG	22	0.172
4	rs2814778	tCTCATTAGTCCTTGGCTCTTA	23	0.230
5	rs917118	tctctTGACTGAGGTCAACGAGC	24	0.230
6	rs1024116	cttGTTCTAATAAAAAGGATTGCTCAT	28	0.241
7	rs7897550	tctctTGTGCAGGATTGAAATATAATT	28	0.115
8	rs722098	agtctgacaaTGACAGTAATGAAATATCCTTG	33	0.460
9	rs10843344	tctctctAGTACTTTGCCAAAGAAACTAAA	33	0.144
10	rs239031	tctctctctctctcAATCTCAGCTTCCACTC	36	0.115
11	rs12913832	ctctctctctctCCAGTTTCATTTGAGCATTAA	37	0.345
12	rs1978806	tctctctctctctcGCAGGAAAGTTGTATTCTGATA	41	0.230
13	rs2040411	ctctctctctctcCTCTGTATTTTCTTACTCTAAGTGC	41	0.115
14	rs773658	ctctctctctctctctctcGGGAAGAATAGAGTCAATCAA	45	0.345
15	rs10141763	tctctctctctctctctcGTGTGAGTTGTGTGATAATCTA	45	0.132
16	rs182549	ctctctctctctctctctctctctctcAGGTGTGAGCCACCG	49	0.747
17	rs1573020	ctctctctctctctctctctctGAGTATTGCCAGCCTGATTC	49	0.144
18	rs896788	ctctctctctctctctctctctctcttctACAGTCACCAGCCAC	53	0.144
19	rs2065160	tctctctctctctctctctctctGCATAGTCTGTTTACTTCATTTG	53	0.149
20	rs2572307	tctctctctctctctctctctctctcatcATTCAATCAATAGTCATAAAC	57	0.126
21	rs2303798	tctctctctctctctctctctctctctctctcAGAAGAGGGACGTGGG	57	0.402
22	rs2065982	ctctctctctctctctctctctctctctGGAAAAAAAAGTCCTCTTTGGTAT	61	0.155
23	rs3785181	tctctctctctctctctctctctctctctctctAGGGCATCTTATCTTGAGC	61	0.086
24	rs881929	ctctctctctctaaactaggtgccacgtcgtgaaagtctgacaaGCCTCTTGCCAGCTCTG	63	0.287
25	rs1498444	tctctctctctctctctctctctctctctctctcGGTTGTCAGAATATTTGCTACA	65	0.345
26	rs1426654	ctctctctctctctctctctctctctctctctctctctttCGCTGCCATGAAAGTTG	68	0.460
27	rs2026721	tctctctctctctctctctctctctctctctctctAAATGATTGACATAGTAGGCTATTG	69	0.201
28	rs4540055	ctctctctctctctctctctctctctctctctctctctctctctcGAAGCAGTGATCAGCAC	73	0.126
29	rs16891982	ctctctctctctctctctctctctctctctctctctctctctctctGGTTGGATGTTGGGGCTT	77	0.201
30	rs1335873	tctctctctctctctctctctctctctctctctctctctctAGGTACCTAGCTATGTACTCAGTAT	77	0.184
31	rs1886510	ctctctctctctctctctctctctctctctctctctctctctctACAAGATTTTCACAACAACACTTGC	80	0.115
32	rs730570	tctctctctctctctctctctctctctctctctctctctctcttCATTAATCACACAAATTTTGCAT	80	0.253
33	rs5030240	ctctctctctctctctctctctctctctctctctctctctctctctctctctcCACAGGAGTGAGCCACTGC	85	0.230
34	rs727811	ctctctctctctctctctctctctctctctctctctctctctctctctctctctCCGGAACTTCAACGACTTA	88	0.287

The additional South Asian specific SNPs were genotyped using pyrosequencing. The sequence context of the SNP was copied from dbSNP into the assay design software, which was used to generate primer sequences in the manner previously described in section 2.5.2. The primers produced are listed below in Table 2.5.

Table 2.5 Primer sequences of extra South Asian/Caucasian divergent SNPs

OligoName	OligoSequence
rs3844336 PCR F	GAGTAGAAACACGGAAAGGTAGCC
rs3844336 PCR R*	GGTGGGGCTGTCTAATCTTG
rs3844336 PyroSeq	GAGTCCTGACCAAATTC
rs602662 PCR F	TGGTGTCGGGAGAACATTG
rs602662 PCR R*	ATGCCATCGCCAGCAAAC
rs602662 Pyro Seq	CATTGACACCTCCCAC
rs9472793 PCR F	CCGGGCAACAGGAAAGAG
rs9472793 PCR R*	CTGGCCACTCAGCTAGGGTACAA
rs9472793 Pyro Seq	GGCAAAGGTTCTCAGT
rs16891982 PCR F*	TCAAATCCAAGTTGTGCTAGACC
rs16891982 PCR R	TCATCTACGAAAGAGGAGTCGA
rs16891982 PyroSeq	GGTTGGATGTTGGGG
Rs35395 PCR F*	TCTCCTGGAACACAAGAATCAGA
Rs35395 PCR R	GTTGCCTACTTGTCATTTCCAGAT
Rs35395 PyroSeq	CCAATAATTATATGAATAGA
Rs2715883PCR F	GTTGGTATTGCTGAAAGCAGTCTT
Rs2715883 PCR R*	TAAACTGGAAATCCGCTTTTCT
Rs2715883 PyroSeq	CATATCTTTGCGCAATT
rs1375131 PCR F	GTCAATGCTCTTGAGCTCTTTACA
rs1375131 PCR R	TGAGAGACAGGCTAATAATATGCA
rs1375131 PyroSeq	TCAAACATCAAAAATAAATT
rs1541317 PCR F*	CCACAGCTTTACTTGCCTGAGGTT
rs1541317 PCR R	CGGCCATGGATATGTGAGAGTTTA
rs1541317 PyroSeq	CCTGGATCTCACTGTTAAAT
rs911903 PCR F*	CGCAGCCATTCTGGAGTTT
rs911903 PCR R	CATTTGAAGCCTCCTGGTATG
rs911903 PyroSeq	TTTTCCTCTCCAGCTC
rs609177 PCR F	ATGTATGTGTACCCATGTGACAGG
rs609177 PCR R*	GATTTCAGGGAGCATGGTGTC
rs609177 PyroSeq	CAGGCTAGAAATCCCTC
rs1040934 PCR F*	CTTCAGTTCCACATCCTAAGTCAC
rs1040934 PCR R	TAGGTAAGCCCCTGAAGATCC
rs1040934 PyroSeq	CCCTGAAGATCCATGC

* denotes that the primer is 5' Biotin labelled.

PCR and pyrosequencing were carried out as detailed in section 2.5.2.

2.9.3 Structure Analysis

Structure is a computerised implementation of a model-based clustering algorithm [42] that is widely used to discern genetic structure within sets of data. The algorithm attempts to look for *K* different genetic signatures within a dataset comprising *n* individuals. Individuals can then either be assigned to one of these *K* clusters using a non-admixture model, or alternatively each individual can be broken down by the proportion of ancestry they have from each of the inferred *K* populations (e.g. a person with mixed ancestry should show components from both/multiple ancestral populations). When the results from the admixture model are displayed graphically, each individual is represented by a single vertical line, and the membership proportion to each inferred *K* group is represented by splitting the line into different coloured segments (e.g. see Figure 1.4 in introduction).

The model uses a Bayesian approach to discern *K* genetic clusters within the data. It does this through the use of allele frequencies. Allele frequencies are assumed to be in Hardy-Weinberg equilibrium and different markers assumed to be in linkage equilibrium within sub-populations: the model partitions the data into sub-populations in order to maximise the generation of appropriate allele frequency distributions across the entire set of markers (e.g. in chapter 5 it can be seen that the genotypes between the Caucasian and Chinese populations for rs3827760 are markedly different, so the model would be unlikely to cluster all the individuals from these two populations in a single *K* group since the genotypes would be exclusively either AA or GG, while Hardy-Weinberg would specify that if these individuals all came from one genetically homogenious population then there should be many more AG genotypes within this group, hence how the model discovers genetic structure within the dataset). The model is not provided with any known information about which population individuals self-identify with, and partitions the dataset purely through the use of this Bayesian model.

In the more complex admixture setting, the algorithm also estimates a parameter *Q* for each individual that specifies how admixed that individual is expected to be (i.e. from

the calculated allele frequencies for each *K* cluster it can estimate the proportion of ancestry within each individual derived from each cluster).

Version 2.3.4 of Structure was used throughout this research, downloadable from http://pritch.bsd.uchicago.edu/structure/release_versions/v2.3.4/html/structure.html. Data was formatted in the manner specified in the accompanying documentation to the Structure program. Briefly, in an excel spreadsheet column A contained the sample number, column B the population identifier (i.e. 1 for Caucasian, 2 for Afro-Caribbean etc., this was used for the subsequent graphical display of the data rather than in the model), column C contained the first allele for the first SNP marker while column D contained the second allele for the first marker, columns E and F contained data for marker 2 etc. Alleles were represented as numbers rather than letters, i.e. allele A was represented with a 1, allele C with a 2, allele G with a 3, allele T with a 4 and any missing data was represented by -9. Row 1 contained a list of all marker names, while rows 2 onward each contained the genotype data for a single sample. The spreadsheet was saved as a 'Tab delimited text file', and this file was then checked in word for the presence of any extra characters (e.g. extra spaces or tabs). An example of such as file is included in Appendix VII.

A new project was created within the structure program and the previously prepared data file uploaded. When prompted information was entered regarding the number of individuals and markers present in the data, and options were ticked specifying that the data included the marker names in row 1 and that the data file stores data for individuals in a single row. Options were also checked to show that individual sample IDs were entered in column A and that the putative population of origin was entered in column B. The parameters set for the model directed a Brunin length of 100,000 and number of repeats of 100,000 using the Admixture Model and independent allele frequencies. *K* values were set depending on the data being analysed.

The data output was displayed graphically using the *Distruct* v1.1 software program [187] downloaded from http://www.stanford.edu/group/rosenberglab/distructDownload.html. Using the output results from *Structure*, the population and individual *Q* matrices were copied across (respectively into the files casia.popq and casia.indivq), the self-

declared population names recorded in casia.names, the parameter file (drawparams) definitions set with the relevant values of *K,* total sample number and number of individual markers used, and the output colours specified (casia.perm).

2.9.4 Snipper App

Population of origin classification was carried out using the Snipper Application [186] as implemented at http://mathgene.usc.es/snipper/. Training set data was formatted in line with the 'example file' downloaded from the website. This entailed entering the genotype data into an excel file such that cell A1 corresponded to the total number of samples in all training sets, cell B1 corresponded to the total number of SNPs tested, and cell C1 corresponded to the number of different populations tested (i.e. number of different training sets). Genotypes were then listed in rows from row 6 down. Column A gave an arbitrary number for every sample in a training set starting at 1, column B listed the population, column C the sample number and columns D onwards the genotypes for each SNP, listed as NN if not determined.

Assessment of classification success with this training set data was achieved by selecting the option 'Thorough analysis of population data of a custom excel file', choosing the previously created excel file and then selecting 'Perform a verbose cross-validation analysis of my population data with the best SNPs'.

Alternatively, the option 'Classification with a custom Excel file of populations (xlsx format)' was selected when classifying new samples using these training sets. The created excel file was selected and the new sample was tested against these training sets by entering a string of letters corresponding to the SNP genotypes in the same order as listed in the excel file, i.e. if analysing 34 SNPs then this string of letters would be 68 characters long (2 alleles per SNP).

2.9.5 Linkage Estimation

Recombination rates vary widely across the genome, but using HapMap data the genetic distance between two markers can be calculated, and this in turn can be converted with a Haldane or Kosambi correction into the chance that a crossing over

event will separate these two markers in a single generation. A genetic distance of 1 centimorgan (cM) is equivalent to a 1% chance of a crossover event, but when looking at more distantly separated loci there is an additional chance of multiple crossing over events 're-linking' the markers, hence the relationship between cM and observed recombination frequency is not so closely observed at higher cM values and requires the correction function.

It is possible to find multiple different reference locations for the same locus, which can make pinpointing the distance between different markers problematic. In this research, locations were taken from the Primary Assembly (alternatively called GRCh37.p2 reference assembly – this is version 37) *via* dbSNP (www.ncbi.nlm.nih.gov/projects/SNP/). Chromosome specific recombination rates can be downloaded from HapMap at: http://hapmap.ncbi.nlm.nih.gov/downloads/recombination/latest/rates/. By finding the difference between the cumulative cM values using the nearest available HapMap SNP to the actual SNP location, a cM value can be ascertained for the distance between two markers. By using the Kosambi formula as implemented in supplementary file 2 of our paper [188], a recombination rate can be determined (for anything under 20cM this value will be nearly identical to the cM value).

2.9.6 Hardy-Weinberg

The genotype results for all individual SNP markers were assessed to check that they were in agreement with Hardy-Weinberg equilibrium – this essentially means that the expected number of homozygote and heterozygote genotypes were present. Allele frequencies were worked out by counting the number of observed alleles and dividing by the total allele number. Expected genotype frequencies (e.g. for AA, AG and GG) can be calculated from the allele frequencies using the formula $p^2 + 2pq + q^2 = 1$ where p represents allele 1 and q corresponds to allele 2. The correlation between the expected genotype frequencies and the observed genotype frequencies can then be tested with a Chi Squared test. An exact test can be used to assess statistical significance between allele frequencies in two different populations.

3 Y Chromosome Results and Discussion

3.1 Development of robust Y-STR multiplexes

During development of the Y-STR suite of multiplex reactions in this project, commercially available Y chromosome kits containing these loci were still 2 years away. It was important to develop a set of reactions that were both sensitive (amplification from <1 ng DNA) and robust, but additionally produced balanced profiles clear of artifactual peaks to enable easy detection of mixtures or contamination. Adherence to these design parameters allowed the reactions to be used in both research situations and (after going through validation procedures) for results produced with these reactions to be admitted legally in forensic casework. For this reason, a number of different PCR optimisation experiments were carried out in order to obtain the desired clear and clean profiles. This included altering the primer concentrations, changing the distribution of loci between the multiplex reactions, altering the magnesium and *Taq* concentrations, and changing the PCR cycling conditions in terms of cycle number, adding an extra elongation step at the end to reduce N-1 (split) peaks and introducing the use of a touchdown PCR methodology. Figures 3.1 – 3.3 show the profiles produced after the optimisation was completed from the two new multiplexes reactions plus the established pentaplex combination.

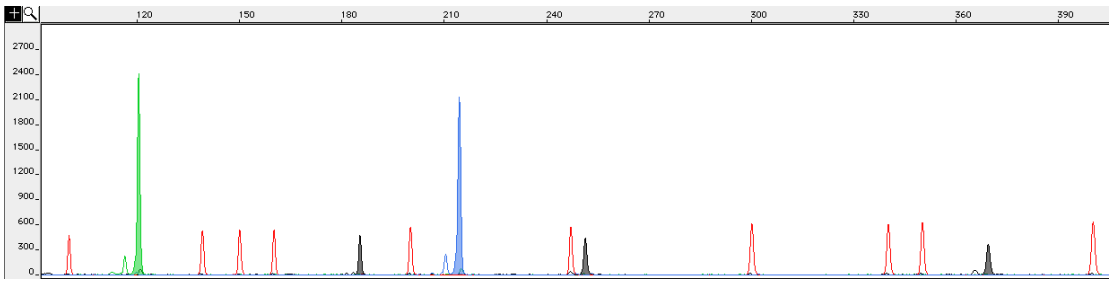


Figure 3.1 Pentaplex Y-STR Amplification
The five loci amplified are DYS19 (Green), DYS393 (Yellow), DYS390 (Blue), DYS389I (Yellow) and DYS389II (Yellow). Products fluorescing yellow are displayed black, while the red data in an internal size standard to allow accurate determination of product length.

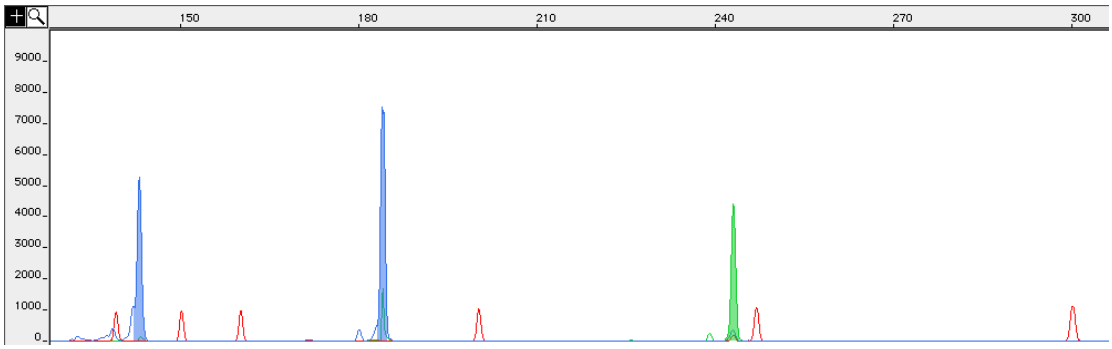


Figure 3.2 Triplex 1 Y-STR Amplification
The three loci amplified are DYS391 (Blue), DYS437 (Blue) and DYS439 (Green). The red peaks represent the internal size standard.

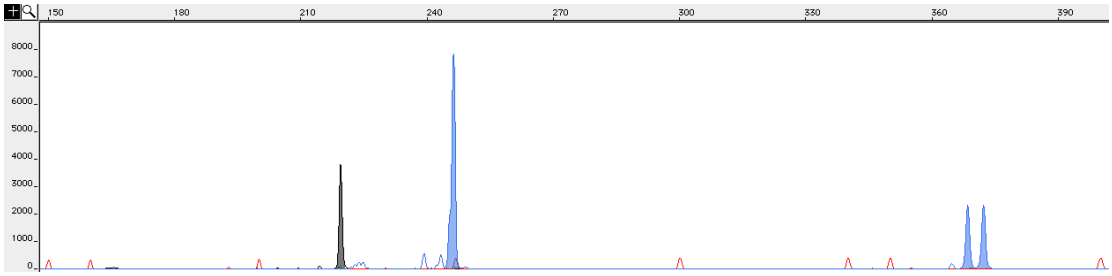


Figure 3.3 Triplex II Y-STR Amplification
The three loci amplified are DYS438 (Yellow), DYS392 (Blue) and DYS385 (Blue). There are 2 copies of the DYS385 locus on the Y chromosome, hence there can be 2 different alleles producing 2 peaks as shown here. The red peaks represent the internal size standard.

Final amplification conditions for the 2 new triplex multiplex reactions are detailed below. Both sets of reactions used final concentrations of 1x AmpliTaq Gold Buffer (Applied Biosystems), 200 µM each dNTP (Promega), and 1.5 mM $MgCl_2$ (Applied Biosystems). 0.5 U AmpliTaq Gold polymerase (Applied Biosystems) was added to each 10 µl reaction. Specific conditions were:

Triplex 1: DYS391, DYS437, and DYS439 were amplified using steps: 95°C 10 min followed by a touchdown PCR with 8 cycles commencing with 94°C 1 min, 60°C 1 min, 72°C 1 min, each cycle reducing the annealing temperature by 0.5°C. This was followed by steps: 94°C 1 min, 56°C 1 min, 72°C 1 min for 22 cycles and a final step of 72°C for 60 min. Primer concentrations were DYS391 0.25 µM, DYS437 0.4 µM and DYS439 0.25 µM, amplifying 1 ng of DNA.

Triplex 2: DYS385, DYS392, and DYS438 were amplified using the same conditions as above with a modification from 22 to 30 cycles for the final steps. Primer

concentrations were DYS385 0.2 μM, DYS392 0.5 μM and DYS438 0.3 μM, amplifying 2 ng of DNA.

Allelic ladders were also generated and sequenced for each locus in accordance with guidelines laid down by the International Society of Forensic Genetics [189, 190]. Sequencing allows the repeat sequence(s) and number to be determined. As an example, figure 3.4 shows the allelic ladder produced for the locus DYS437 - all components were shown to be free of sequence anomalies and correspond to alleles 7, 8, 9, 10 and 11.

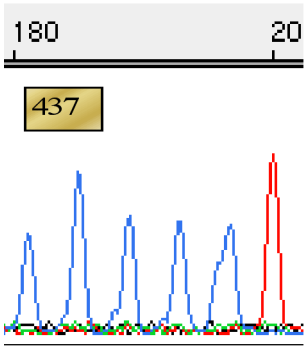


Figure 3.4 Allelic ladder produced for DYS437
Blue peaks correspond to alleles 4bp apart while the red peak is a size standard of 200bp.

3.2 Characterisation of Markers

If one aim is to use the Y-STR haplotypes to trace back lineages in either kinship cases or population structuring, then it is vital to know how often the markers within that haplotype are likely to change, and what that change is likely to be. It is known that microsatellite markers mutate at different rates, and while a hypermutable marker would increase discrimination it would also decrease the utility of the loci for relationship or population discrimination purposes since there is a greater chance of an allele changing between individuals of the same lineage. At the time of this study there was no mutation information available for DYS437, DYS438 and DYS439, and variable amounts for the other loci.

3.2.1 Mutation Rates

Y-STR mutation rates are easily estimated by observing the transmission of genetic information from father to son. A total of 3049 STR germline transmissions were observed in 250 father:son pairs and 13 mutations were seen, giving an overall Y-STR mutation rate of 4.26×10^{-3}. These mutations occurred in 8 of the 13 Y-STRs studied, and sequencing confirmed that all were shown to result from a change in the number of repeat units within the variable portion of the marker (as opposed to changes in the flanking region). Of the 13 mutational events, all resulted in a 1 step allelic change, and there was no tendency for repeat gain or loss (there were six augmentations and seven diminutions). Details of the individual mutations and overall marker rates are shown in Tables 3.1 and 3.2. In all families presenting with Y-STR mutations, a set of autosomal STR markers were additionally tested to prove paternity (detailed in Appendix I), and in all cases a paternity index of over 1 million was obtained (Table 3.1), proving that the mutations are genuine rather than a consequence of non-paternity.

Table 3.1 Father-Son Mutations Observed

Loci	Father	Son	Population Group	Paternity Index
DYS385	13-14	13-15	Danish Caucasian	67,325,823
	11-14	10-14	Irish Caucasian	12,106,373
	15-17	15-18	British Afro-Caribbean	6,899,322,090
	11-15	12-15	British Caucasian	184,346,952*
DYS389II	29	30	Irish Caucasian	618,164,379
	29	30	British Caucasian	4,393,947*
DYS391	11	10	Danish Caucasian	55,911,395
	10	11	Albanian	6,592,914
DYS19	16	15	Polish Caucasian	6,644,615
DYS389I	15	14	British Caucasian	26,475,036
DYS437	16	17	British Caucasian	10,113,112
DYS439	11	10	Jersey Caucasian	4,948,446
DYS460	11	10	British Afro-Caribbean	176,442,573

* Additional tests required to achieve this value.

Table 3.2 Mutation rates for 13 Y chromosome STRs

DYS Loci	No. Of Mutations	No. Of meioses studied	Mutation Rate (95% CI) x 10^{-3}	Combined[§] Mutations / Meioses	Combined[§] Mutation Rate (95% CI) x 10^{-3}
DYS19	1	245	4.08 (0.2–26)	22/9049	2.43 (1.5-3.7)
DYS385[#]	4	472	8.45 (0.3–23)	30/13678	2.19 (1.5-3.2)
DYS389I	1	247	4.05 (0.2–26)	17/7253	2.34 (1.4-3.8)
DYS389II[*]	2	246	8.13 (1-32)	25/7240	3.45 (2.3-5.2)
DYS390	0	248		19/8531	2.22 (1.4-3.5)
DYS391	2	248	8.06 (1–32)	25/8480	2.95 (1.9-4.4)
DYS392	0	226		4/8444	0.47 (0.2-1.3)
DYS393	0	248		6/7233	0.83 (0.3-1.9)
DYS437	1	249	4.02 (0.2–26)	5/3237	1.54 (0.5-3.8)
DYS438	0	225		2/4209	0.48 (0.1-2)
DYS439	1	249	4.02 (0.2–26)	26/4253	6.21 (4.1-9.1)
DYS460	1	74	13.5 (0.7–83)	5/1109	4.51 (1.7-11)
GATA A10	0	72		4/946	4.23 (1.3-12)
TOTAL	13	3049	4.26 (2.2–7.5)	190/83662	2.27 (2.0-2.7)

[*] Mutations in DYS389I are not included here

[#] As there are 2 copies of the DYS385 locus on the chromosome, this rate represents the chance of mutation per locus rather than for the combined DYS385 marker.

[§] Combined data from this study plus published Y-STR mutation investigations [191-201].

The overall Y-STR mutation rate of 4.26 x 10^{-3} is slightly higher than other estimates of 2.8 x 10^{-3} [202] and 2.3 x 10^{-3} [195], but these are still within the 95% confidence limits of the value determined here. Due to the rare nature of mutational events, the locus specific mutation rates are all associated with large confidence intervals and it is impossible to draw any statistically significant conclusions regarding individual mutation rates. By combining this data with other published studies [191-201] it is possible to improve the accuracy of the mutation rate estimates, as can be seen in Table 3.2. Significant differences can now be observed between some loci, for example DYS392, DYS393 and DYS438 all have significantly lower mutation rates than the other loci while DYS439 has a significantly higher mutation rate than virtually all other Y-STR loci studied. This information is vital if using these markers for relationship casework, but also has implications in lineage/population genetics studies since changes in the slowly mutating markers are likely to be of more

significance and more likely to highlight population structure since they are more stable and hence less prone to further mutation.

3.2.2 Non-standard Allelic Patterns

A number of unusual allelic states were observed in various samples throughout the course of this research; the cause of these also being non-standard transmission of the Y chromosome at some stage of the individual's paternal line. Seven intermediate alleles were observed, six individuals presented with allele duplications and two sample produced profiles containing null alleles. The structure of the intermediate alleles is presented in Table 3.3 and it was determined that three were due to incomplete repeats while four were due to deletions within the flanking region of the STR, both additional types of mutation that must have occurred during meiosis rather than simple polymerase slippage leading to a change in repeat number as observed in the father-son pairs above. Intermediate alleles were seen most often in the DYS385 marker (6/7 instances), and three individuals possessed the same 16.3 allele caused by a T deletion in the STR flanking region. Two of the haplotypes where this allele was seen differed at only 2 other loci, which may suggest that at least two of these three 16.3 samples have a common lineage, although they all originate from different geographical regions: Albania, The Channel Islands and South Asia. In the populations studied, these intermediate alleles were observed rarely, suggesting that these mutation events do not occur commonly.

Table 3.3 Intermediate alleles

A 10.2 allele observed in locus DYS392 in a Danish Caucasian sample due to the deletion of a Thymine in the 5' flanking region as shown below.

Forward Primer ctaatttgatttcaagtgtttgttatttaaaagccaagaaggaaaacaaatttttt[-]cttgtatcaccatttattt(att)$_{11}$tact aaggaatgggattggtaggtttaatgatccctctgttttgacttctttgaga *Reverse Primer*

A 14.2 allele observed in locus DYS385 in an Indian sample due to a 2 base pair deletion or insertion at the end of the repeating section.

Forward Primer gacaccatgccaaacaacaacaaagaaaagaaatgaaattcagaaaggaaggaaggaaggagaaagaaagtaaaaa agaaagaaagagaaaaagagaaaaagaaagaaagagaagaaagagaaagaggaaagagaaagaaaggaaggaaggaaggaaggaagg (gaaa)$_{14}$gaaagaaaaagaaaggaggactatgtaattggaatagatagattatttttaaaatatttttattacctttacagttttttaaatgccgccatttca gaaagaaatctggtcagcagcccttac *Reverse Primer*

A 13.2 allele observed in locus DYS385 in an American Caucasian sample due to an irregular repeat unit in the middle of the repeat structure.

Forward Primer gacaccatgccaaacaacaacaaagaaaagaaatgaaattcagaaaggaaggaaggaaggagaaagaaag taaaaaagaaagaaagagaaaaagagaaaaagaaagaaagagaagaaagagaaagaggaaagagaaagaaaggaaggaaggaaggaagga agg(gaaa)$_{6}$aa(gaaa)$_{7}$gagaaaaagaaaggaggactatgtaattggaatagatagattatttttaaaatatttttattacctttacagttttttaaatgc cgccatttcagaaagaaatctggtcagcagcccttac *Reverse Primer*

A 16.2 allele observed in locus DYS385 in a British Afro-Caribbean individual due to an irregular repeat unit near the beginning of the repeat structure.

Forward Primer gacaccatgccaaacaacaacaaagaaaagaaatgaaattcagaaaggaaggaaggaaggagaaa gaaagtaaaaaagaaagaaagagaaaaagagaaaaagaaagaaagagaagaaagagaaaggaaagagaaagaaaggaaggaaggaagg aaggaagg(gaaa)$_{1}$ga(gaaa)$_{15}$gagaaaaagaaaggaggactatgtaattggaatagatagattatttttaaaatatttttattacctttacagttttttt aaatgccgccatttcagaaagaaatctggtcagcagcccttac *Reverse Primer*

Three 16.3 alleles observed in the DYS385 locus. One in an Albanian sample, one in a South Asian sample and one in an individual from Jersey. All had the same deletion of a Thymidine in the poly-T stretch found in the 3' flanking region.

Forward Primer gacaccatgccaaacaacaacaaagaaaagaaatgaaattcagaaaggaaggaaggaaggagaaagaaag taaaaaagaaagaaagagaaaaagagaaaaagaaagaaagagaagaaagagaaaggaaagagaaagaaaggaaggaaggaaggaagga agg(gaaa)$_{17}$gagaaaaagaaaggaggactatgtaattggaatagatagattatttttaaaatatttttattacctttacagttttt[-]aaatgccgccatt tcagaaagaaatctggtcagcagcccttac *Reverse Primer*

Events leading to both allele duplication and allele deletion were also observed, at least some of these being caused by structural chromosomal changes encompassing many thousands of DNA bases. Of the deletion mutations, one case involved the loss of the DYS390 allele which was not amplified even when an alternate set of primers were designed - suggesting that a primer binding site mutation was not the cause of the null allele. In the other case DYS385, DYS392 and DYS460 were absent, markers situated in close proximity to each other, suggesting that an area of the Y chromosome had been deleted in that individual that spanned at least 1.8Mb. These three STRs reside in the azoospermia factor b (AZFb) region, and deletions in this area are known to cause severe spermatogenic defects [203].

Of the duplications observed, two individuals had multiple copies of DYS19, two multiple copies of DYS385, and two presented with duplications in both DYS389 and DYS439. DYS389 and DYS439 are neighbouring markers on the chromosome, and a mechanism for explaining duplication events affecting both of these markers has been previously postulated [204]. When these rare duplication events are found, they can be very useful ancestry markers highlighting population structure. For example a duplication at DYS19 is known to have occurred within the lineage defined by SNP M48 (haplogroup C3c [205]), and hence be at a high frequency in Central Asia (especially Kazakhstan) where this haplogroup is found. The allelic state of M48 was tested in both samples displaying DYS19 duplications, but the substitution was not present in either individual suggesting that the duplication had arisen independently of that present in C3c. Another example of the power of rare duplication events to identify population structure is detailed in Figure 3.5 below. The case described shows how it was possible for us to reconstruct a family tree extending over 16 generations by testing the Y chromosome. Four male individuals were shown to have directly descended from a common ancestor from the 18th Century by sharing the same 11 Y-STR haplotype, including the rare duplications at DYS389I (13-14), DYS389II (29-30) and DYS439 (11-13).

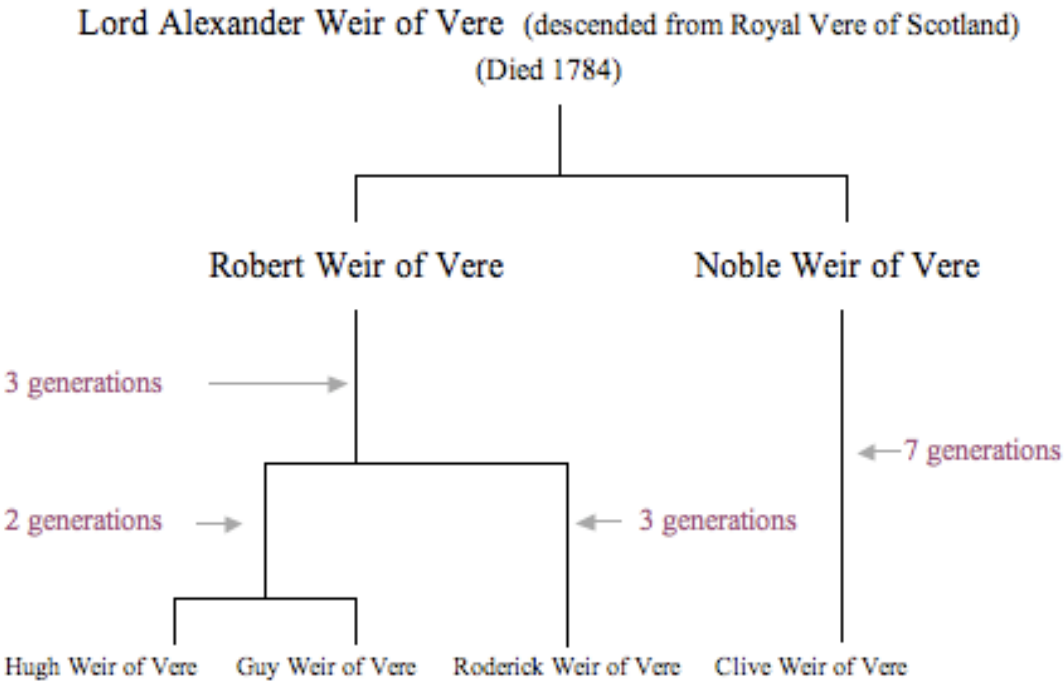


Figure 3.5 Male lineage from an extended family study undertaken at our laboratory
Four of the individuals typed share the same rare haplotype [including DYS389I 13,14 ; DYS389II 29,30 and DYS439 11,13] and are listed at the bottom of the tree with the common ancestor shown at the top.

3.3 Caucasian Samples

Complete 11 marker Y-STR haplotypes were produced for 250 British Caucasian samples. There were 220 different haplotypes produced from the 250 samples (88%) and these can be found within pages 293-296 of the paper included in Appendix II. An additional 155 Irish samples were typed to enable a comparison to be made between the British individuals and a geographically close population known to have strong Celtic influences. A total of 127 different haplotypes were coded for by these 155 Irish samples (82%) and they can be found in pages 66-68 of the paper included in Appendix III. Twenty-six haplotypes were present in both populations, accounting for 90 of the 405 samples tested, showing a not unexpected degree of genetic similarity between the Irish and British Caucasian populations.

When comparing these two sample sets, the most notable difference is in the respective populations' variability. Allele distributions are similar for all markers, however the marker diversity values (an indicator of polymorphism) are lower for all

markers in the Irish population compared to the British Caucasian population, as shown in Table 3.4.

Table 3.4 Locus Diversity Value in the Irish and British Caucasian Populations

Marker	Irish Diversity	UK Caucasian Diversity
DYS385	0.764±0.028	0.807±0.023
DYS390	0.664±0.026	0.706±0.017
DYS439	0.634±0.028	0.650±0.017
DYS389II	0.609±0.037	0.707±0.017
DYS391	0.510±0.022	0.514±0.008
DYS392	0.505±0.044	0.562±0.028
DYS389I	0.459±0.042	0.522±0.030
DYS437	0.393±0.044	0.540±0.027
DYS19	0.327±0.043	0.466±0.033
DYS438	0.311±0.044	0.547±0.029
DYS393	0.235±0.044	0.323±0.037

The decrease in Irish variation is underlined by a lower haplotype diversity value for the Irish sample (0.99598±0.00167 versus 0.99838±0.00062), suggesting that the population may have a more conserved degree of patrilineage within a smaller geographical area. This reduced variability and genetic independence is reflected when comparing the Irish population samples with the 26,434 samples from 234 populations contained within the global Y Haplotype Reference Database (YHRD, www.yhrd.org release 23). There are eight instances, accounting for 12.3% of samples, where a 10-marker haplotype (excluding DYS437) is observed multiple times within the small sampled Irish population (between 2-4 times, a frequency of 0.013-0.026) but has either not been seen at all in the other 234 populations, or has only been seen outside Ireland a handful of times (a worldwide frequency of <0.0005). This further suggests that, at the very least, a significant proportion of the current Irish population descended from a relatively small gene pool evolving distinctly from the rest of the world. In comparison, while a small subset of duplicated (observed 2-3 times) British Caucasian haplotypes (6.8% of samples) are also found rarely worldwide (frequency of <0.0005), they are all observed a minimum of 3 times outside the UK, highlighting the increased variation in the British sample as well as the gene flow with neighbouring populations. The most commonly observed haplotype within Europe (the same 10 marker profile is seen in over 200 samples) is also the most frequently observed haplotype within both population

samples, but symptomatic of the reduced Irish variety is found at a frequency of 4.5% in the Irish population sample and at 2.8% in the UK one.

A recent study on haplotype sharing within surnames (since surnames and the Y chromosome are both inherited patriliniarly, there is some correlation between particular surnames and specific Y-STR haplotypes) has also found that there is a greater degree of shared ancestry in the Irish population. Even relatively common Irish surnames (>40,000 individuals) show a high degree of co-ancestry, while these patterns can only be discerned in much rarer surnames within the UK (surnames with >10,000 individuals in the UK demonstrate no significant haplotype clustering) [206].

The distribution of haplotypes within the two population groups can also be displayed graphically using Network (as detailed in section 2.7.1): the Irish population represented in Figure 3.6 and the UK Caucasian population in Figure 3.7. Each circle represents a haplotype, while the lines between the circles represent the differences between the haplotypes - the larger the circle or longer the line the more samples/differences there are. Comparing these two figures, it is once again obvious how little variation there is within the samples Irish population.

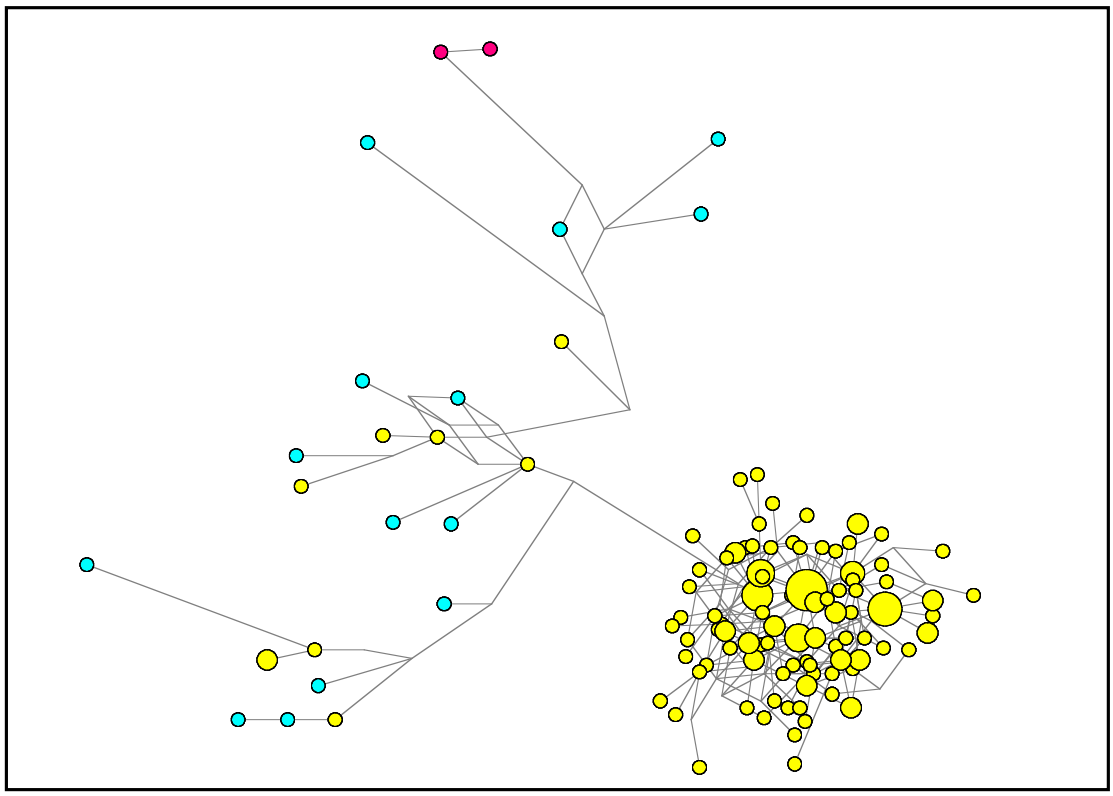


Figure 3.6 Network of Irish Population Samples
The circles represent the samples and the lines the relationship between samples (allelic differences). The larger the circle the more samples with identical haplotypes, the longer the line the more differences between the linked haplotypes. All samples are Irish but, of those not in the central core, those haplotypes labelled in blue are unique amongst all other 26,434 11-marker samples contained in the YHRD, while those haplotypes labelled in pink are predominantly observed in Eastern Europe.

Figure 3.6 demonstrates that the vast majority of Irish samples cluster very tightly, only being different from one another by a few mutational steps. Twenty-six of the 155 samples form a breakaway branching structure, separated from the clustered majority due to characteristics in the (slowly mutating) markers DYS438 and DYS392 (the main partitioning branch depicts a 2-step loss in both markers). These individuals would appear to have evolved separately to the majority of the sampled Irish population: the chances of this being a recent branch from the main Irish cluster are remote considering the main branching point contains 2-step mutations (most likely occurring in two 1-step events) in 2 of the slowly evolving Y-STR markers, as well as many other additional mutations separating the individual represented in this branched structure. Of these 26 samples, 2 show strong Eastern European affiliation (in pink), one possibility being that this reflects recent migration, while most of the remainder are not only genetically distinct from the clustered Irish samples but also of

the worldwide population. Fourteen of these remaining 24 individuals are only observed in this Irish population sample and are not seen at all in 26,434 other worldwide samples – again a trait that may be related to Ireland being an island, but is also probably associated with the current coverage of the Y-STR database with a lack of samples from locations like Boston that experienced high levels of Irish immigration. Taking all of this data into account, it suggests that most of these 26 separated samples most probably originate from individuals split from the main cluster a long time ago and who evolved separately in remote or isolated parts of Ireland (a founder effect). Another explanation would be that these different lineages arose independently from individuals with a different genetic background, through either variation in the origins of those in the early Irish settlements or more recent immigration, although if immigration was the cause then it would be unexpected to find so many samples with absolutely no matches throughout the rest of the world. Either case is supported by the lack of intermediate haplotypes between the tight cluster and these singletons.

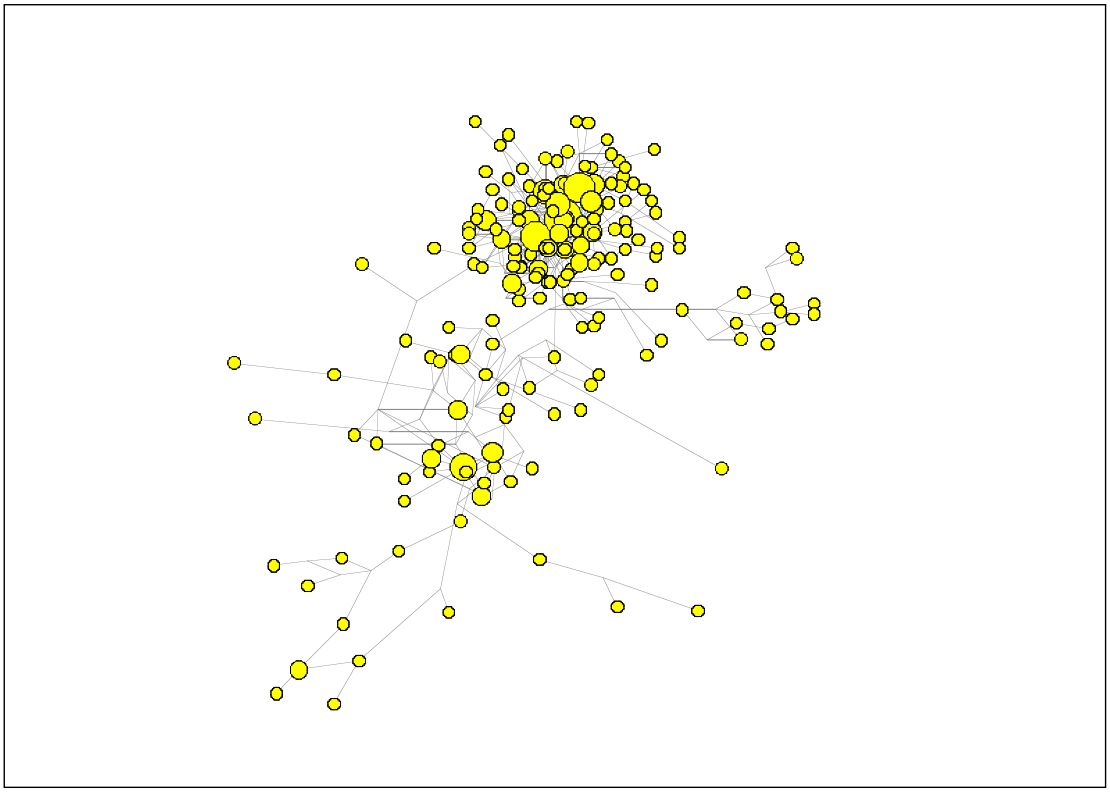


Figure 3.7 Network of British Caucasian samples
Haplotypes are represented by circles – the larger the circle the more samples possess that haplotype. The lines between spheres represent the number of mutational changes between haplotypes – the longer the lines the more distantly related are the haplotypes.

The cluster of haplotypes within the British Caucasian population (present at the top of Figure 3.7) represent broadly similar haplotypes to those clustered together in the Irish population, however it can be seen that there is a much more extensive network extending from this central cluster. Given the many cultural influences on the British gene pool throughout history, combined with a larger population spread over a greater landmass, the higher diversity is not surprising. To assess the similarity of the male genome in British Caucasian individuals to other relevant populations, and possibly infer from that any genetic influences still dominating in Britain, comparisons have been made not only with the Celtic Irish samples, but also with other European populations. To this end, the Forensic Y Chromosome Research Group (see page 289 of [175]), of which I am a contributing member, pooled the data from 69 participating European Laboratories and published the findings [175]. It was possible to amass 12,727 samples from 91 European populations.

The analysis included the calculation of Φst values which provide a measure of the genetic distance between populations, and enabled the grouping together of closely related data sets to give a reduced total of 45 populations/meta-populations. Clustering of population samples was then possible based on minimal Φst values, as shown in Figure 3.8. Not surprisingly, the British Caucasian population falls into a Western Europe cluster including Ireland, Northern Italy (though not Central or Western Italy), Spain and Portugal, South Holland (though not North Holland or Belgium) and Alsace in France (though not the more cosmopolitan Paris). Considering the historic role that Germany played in shaping the British population, the current German populations bear greater similarity to other Central European population and the British population clusters with those geographically to the West of Europe, including those populations known to have strong Celtic influences such are Ireland and parts of Iberia. The inclusion of the Northern Italian region in this Western population cluster rather than with the other Italian populations is unexpected. It may reflect some degree of Roman influence in the British Caucasian population, but even if that is the case there should still be strong similarity with the other geographically proximal populations around Northern Italy. It has also been stated [207] that a proportion of the Roman Legionaries, far from coming from Italy,

would actually have come from the conquered Roman provinces including Belgian Gaul and Germania, as well as contingents of German mercenaries, suggesting that any Roman influence would be weakened if comparing to a modern day Italian population.

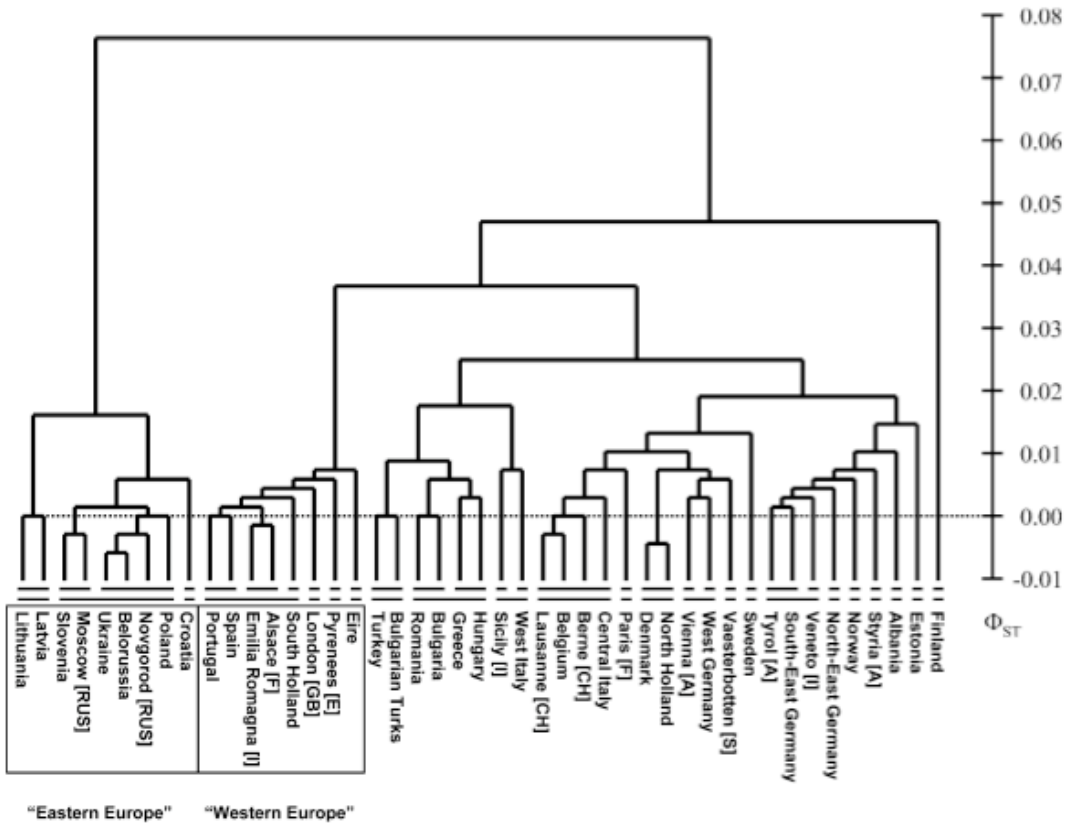


Figure 3.8 Relationship between 45 European populations based on Φst values
Taken from Roewer *et al.* [175]. Submitted British and Irish populations cluster in the Western Europe box.

When the pair-wise Φst values were subjected to a multi-dimensional scaling analysis, it was determined that 96.4% of all Y-STR variation across the continent could be accounted for in 3 dimensions (3 different genetic clines across Europe). These 3 dimensions are graphically portrayed in Figure 3.9. Graph A accounts for the majority (88.7%) of the variance on its own and the 'Western Europe' populations identified in Figure 3.8 are clearly identified on the Western periphery of this East-West stratification of genetic variation. The East-West cline clearly shows three differing areas of Y-STR haplotype distribution: an Eastern zone, a Central zone, and a Western zone. It has been suggested that this represents a major linguistic divide

between Slavic speaking Eastern Europe, and the Latin derived Romance languages predominately used in Western Europe, with a central zone consisting of Italian and Germanic speaking populations [175].

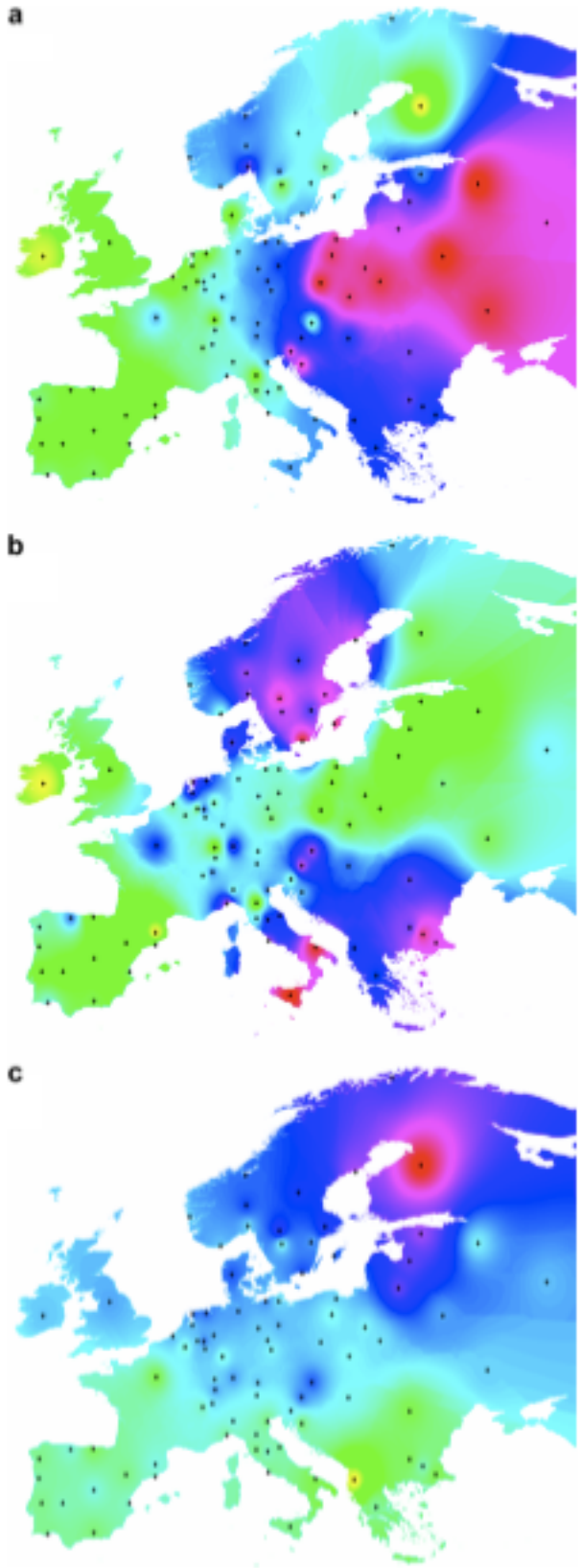


Figure 3.9 Genetic contour maps of Europe depicting 94.6% of Y-STR variation
Taken from Roewer *et al.* [175]. East-West cline in map A represents 88.7% of variation, map B 5.7% and map C 2.2%. Genetic distance is depicted by colour, from yellow to pink through green and blue.

The second dimension (map B in Figure 3.9) explains 5.7% of the Y-STR variation and further highlights the similarity of the North Italian Emilia Romagna population to the UK as well as confirming the tendency of Irish genetic isolation with the population once again being at the extreme end of both the genetic and geographic cline. The conclusion drawn from this comparison of the British Caucasian data with other European populations is of a clear Y-STR genetic similarity to other Western European populations, with only a minor genetic influence shown to be associated on a latitudinal basis (2.2%, graph C of Figure 3.9). The apparent similarity to the North Italian region of Emilia Romagna (but less so to surrounding areas) bears further investigation.

3.4 British Population Y-STR Data

Eleven marker haplotypes were generated from 250 Afro-Caribbean samples and 250 South Asian samples, and are listed in full in Appendix II (on pages 296-304 of the included paper). Diversity was highest in the Afro-Caribbean population set with 241 different haplotypes obtained from the 250 samples (96%), whilst in the South Asian set 227 different haplotypes were produced (91%), and this compares with 220 (88%) in the British Caucasian set. Table 3.5 displays the locus diversity values for each STR marker in the three different populations, and is can be seen that some loci show far more varied allele distributions in particular populations; for example DYS390 exhibits reduced diversity in the Afro-Caribbean population while DYS393 is virtually monomorphic within the Caucasian population (82% of samples possess a 13 allele) despite displaying an appreciable level of polymorphism within the other two populations. Clear allele frequency differences are also discernable in most of the Y-STR loci with the exception of DYS389I and DYS391. Figures 3.10, 3.11, 3.12 and 3.13 show particularly striking differences seen at DYS390, DYS392, DYS438 and DYS385.

Table 3.5 Gene diversity values for 11 Y-STR markers across 3 British populations

Caucasian		Afro-Caribbean		South Asian	
Loci	Diversity	Loci	Diversity	Loci	Diversity
DYS385	0.8078	DYS385	0.9489	DYS385	0.9324
DYS389II	0.7168	DYS19	0.7331	DYS389 II	0.7925
DYS390	0.6999	DYS389II	0.7306	DYS390	0.7728
DYS439	0.6480	DYS393	0.6684	DYS439	0.7286
DYS392	0.5626	DYS439	0.6225	DYS19	0.6754
DYS438	0.5560	DYS390	0.5271	DYS438	0.6715
DYS437	0.5377	DYS389I	0.5095	DYS393	0.6642
DYS391	0.5173	DYS437	0.4311	DYS389 I	0.6165
DYS389I	0.5148	DYS438	0.4226	DYS392	0.5349
DYS19	0.4607	DYS391	0.3186	DYS437	0.4868
DYS393	0.3296	DYS392	0.3044	DYS391	0.3855

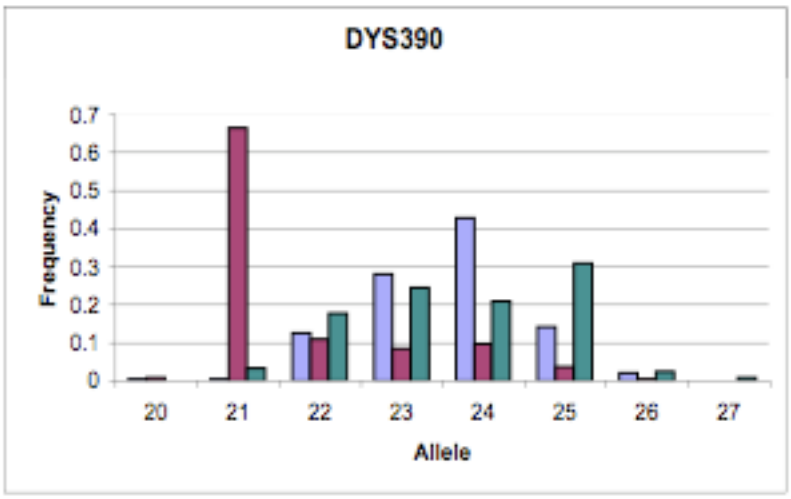


Figure 3.10 DYS390 Allele Frequency Distribution in 3 British Populations
Caucasian frequencies are shown in blue, Afro-Caribbean in red and South Asian in green.

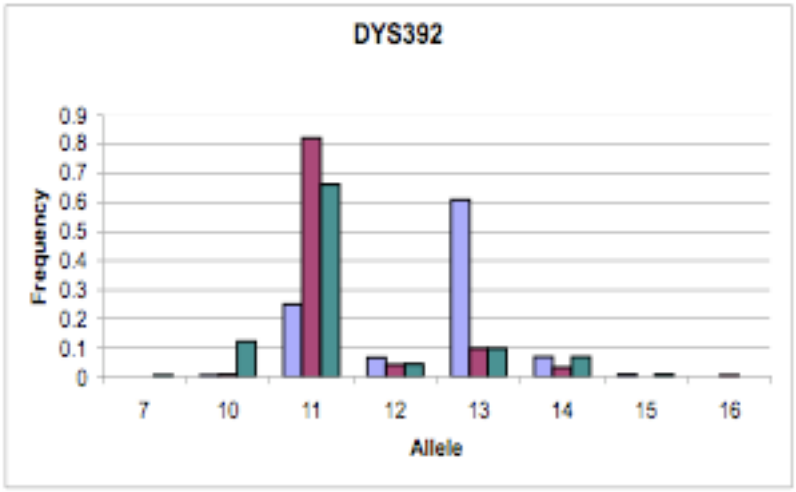


Figure 3.11 DYS392 Allele Frequency Distribution in 3 British Populations
Caucasian frequencies are shown in blue, Afro-Caribbean in red and South Asian in green.

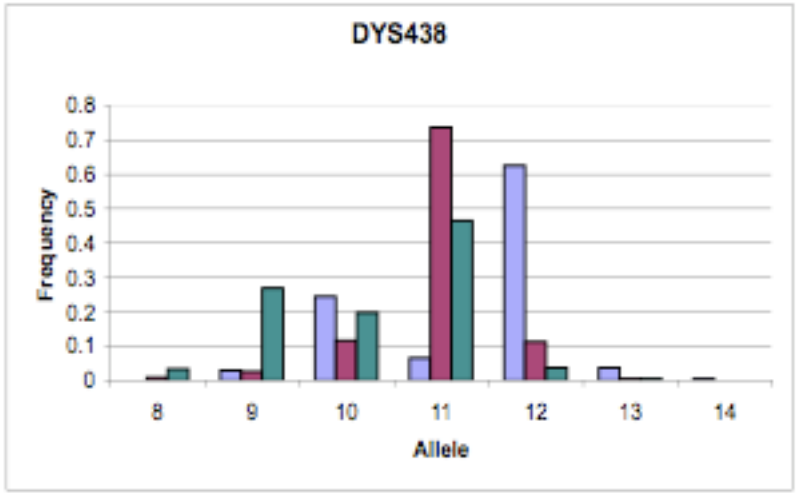


Figure 3.12 DYS438 Allele Frequency Distribution in 3 British Populations
Caucasian frequencies are shown in blue, Afro-Caribbean in red and South Asian in green.

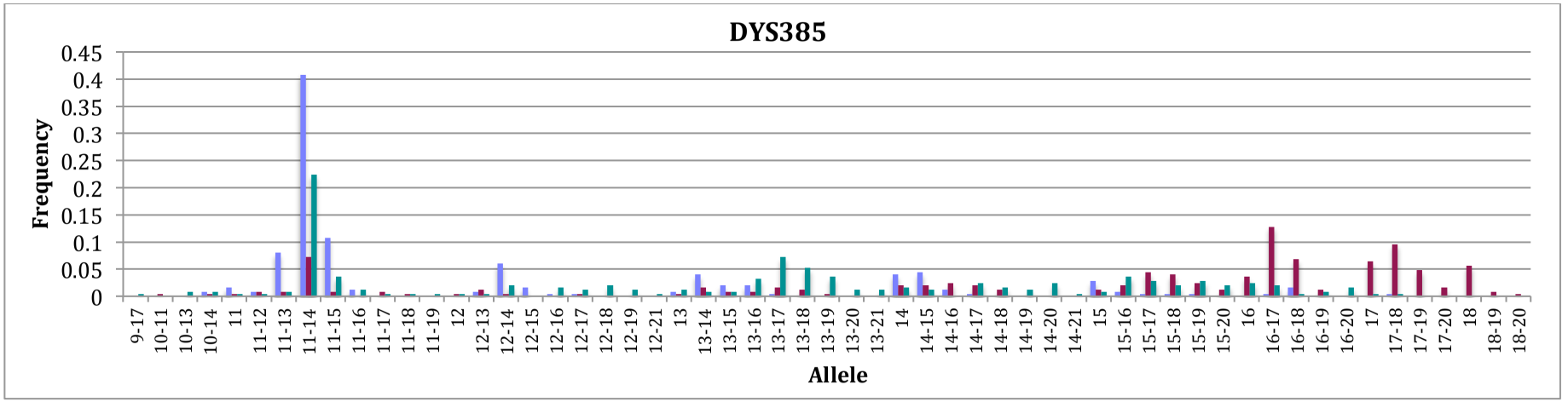


Figure 3.13 DYS385 Allele Frequency Distribution in 3 British Populations
Caucasian frequencies are shown in blue, Afro-Caribbean in red and South Asian in green.

AMOVA calculations were performed on the three main British populations to show the degree of similarity in the haplotype composition between the different sample groups. A population sample from an unrelated European city (Barcelona in Spain [208]) was also included in the comparison to provide a baseline level of variance between two unrelated but genetically similar populations. The results are presented in Table 3.6 as R_{ST} values, the larger the value the more dissimilar the populations are as revealed by the rapidly evolving STR genetic markers on the Y chromosome. The R_{ST} value represents the increase in variance seen when the populations are merged rather than analysed separately, and takes into account not just the number of markers that are different but also the degree of difference (i.e. how many allelic step changes there are). All populations were shown to be statistically different from one another, the p value when comparing the British Caucasian and Barcelona populations was 0.018, while all other p values were less than 0.00001. The two European samples (British Caucasians and Barcelona) are shown to be genetically similar, and over 99% of genetic variation can be seen within the individual populations, with variance only increasing by less than 1% if the two population samples freely mixed together. In contrast, there is a high degree of genetic difference between the three British populations tested, which is to be expected since they represent individuals with ancestry derived from three different continents. The highest degree of genetic difference is seen between the European and African samples and the least between the European and South Asian samples; this is to be expected both in terms of geographic distance and genetic divergence as predicted by the out-of-Africa theory of human evolution.

Table 3.6 AMOVA Pairwise R_{ST} Results

Populations	Pairwise R_{ST} Value
Caucasian : Afro-Caribbean	0.37171
Afro-Caribbean : South Asian	0.27620
Caucasian : South Asian	0.20447
UK Caucasian : Barcelona	0.00962

Despite the pronounced differences at both individual markers and the haplotype level, there are still 10/678 haplotypes that are present in more than one population. These shared haplotypes could have arisen separately within the different populations

(identical by state) or there could be individual movement between the different populations (haplotypes therefore having a shared ancestor and are identical by descent). There are seven Afro-Caribbean haplotypes that are also observed within the UK Caucasian population sample, and from Figure 3.14 it can be seen that most of these shared haplotypes are associated closely with the main Caucasian cluster, and are presenting with haplotypes very different from the majority of other Afro-Caribbean samples – this would seem to indicate that the most likely source of these shared haplotypes is Caucasian Y chromosome penetration into the Afro-Caribbean population at some stage in the past. Two haplotypes were also shared between South Asians and Caucasians, and one between South Asians and Afro-Caribbeans. There is a striking frequency difference at DYS390 (Figure 3.10) between the Afro-Caribbean population and the others, however at most other loci it is the Caucasian samples that are showing a markedly different allele distribution, and this is reflected in Fig 3.14. The main separating branch between the Caucasian cluster and the rest consists of differences at markers DYS438 and DYS393. The co-localisation of the Afro-Caribbean and South Asian samples would seem to be in conflict with the AMOVA analysis of Y-STR haplotypes which shows that overall the South Asian and Caucasian samples share more in common genetically than the South Asian and Afro-Caribbean samples. Increased clarity and separation in the area of the network representing the majority of Afro-Caribbean and South Asian samples is therefore needed to more accurately graphically represent the true relationship between the 11-marker haplotypes of the three British populations.

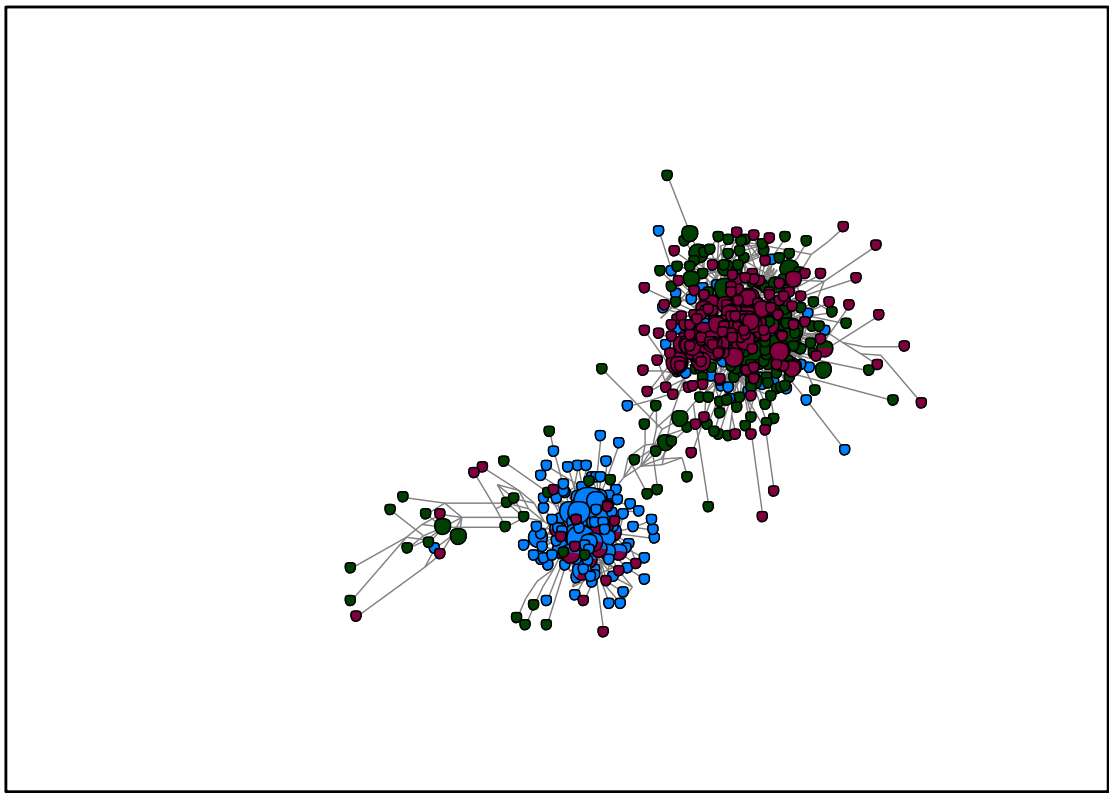


Figure 3.14 Network of 3 British Population Samples - Caucasian, Afro-Caribbean and South Asian
Caucasian haplotypes are shown in blue, Afro-Caribbean in red and South Asian in green. The size of the circle is proportional to the number of samples represented by the haplotype, and the length of the line separating 2 haplotypes is proportional to the number of allelic step changes between the haplotypes.

Figure 3.14 shows the calculated hypothetical relationship between the different samples if changes at all markers are equally weighted, and Figure 3.15 displays only the torso of this network (the main framework to which other samples are added). Figure 3.16 shows how this torso is changed, resulting in a superior differentiation between population groups, when the importance of mutations at individual markers is altered to reflect the mutation rates calculated in section 3.2. In Figure 3.16 the weight of mutations in DYS392, DYS393 and DYS438 is tripled, to reflect the fact that the mutation rate of these markers is at least 3 times lower than most other markers tested, and hence changes at these loci are likely to happen less frequently and have less chance of recurring independently in multiple different lineages. The weight of mutations at DYS439 is halved because this locus has a mutation rate at least double to most other markers, and hence alleles at this marker are less stable. By utilizing the knowledge about the genetic characteristics of these different markers, it is possible to draw more realistic conclusions about how the samples could

be genetically related. This can have implications when designing population determination algorithms to distinguish haplotypes from individuals in the 3 different population groups.

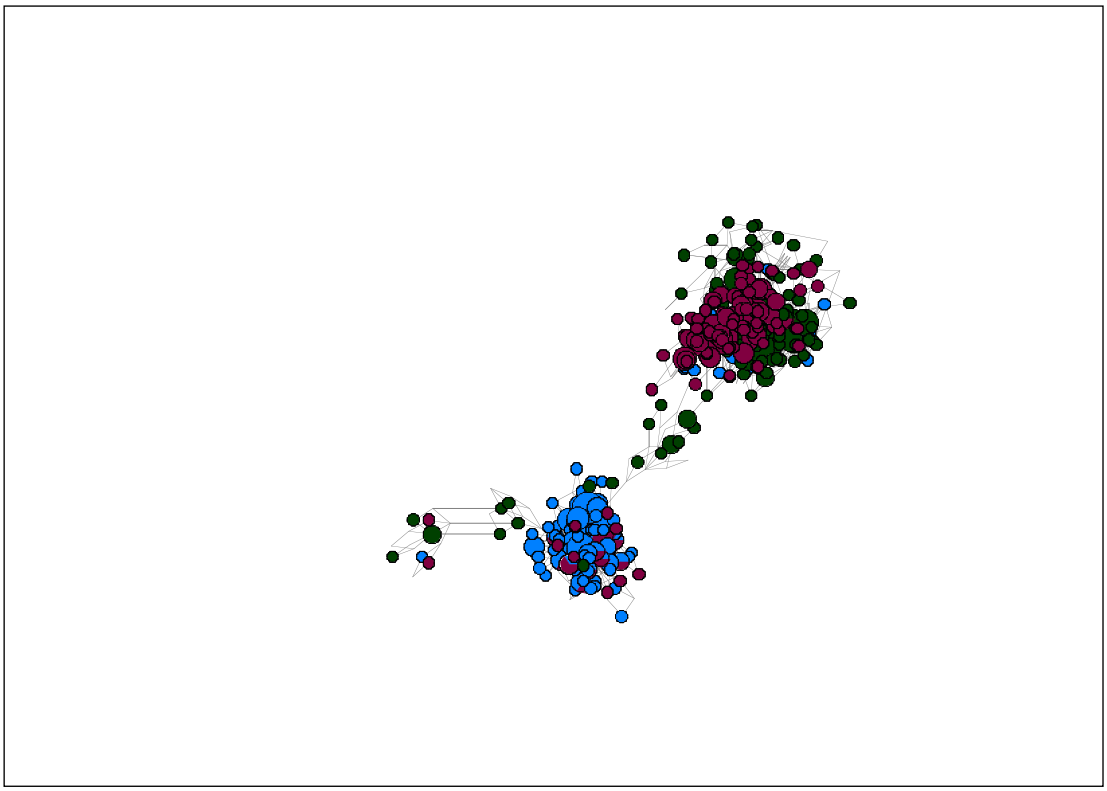


Figure 3.15 Torso of British population network shown in Figure 3.15
Caucasian haplotypes are shown in blue, Afro-Caribbean in red and South Asian in green.

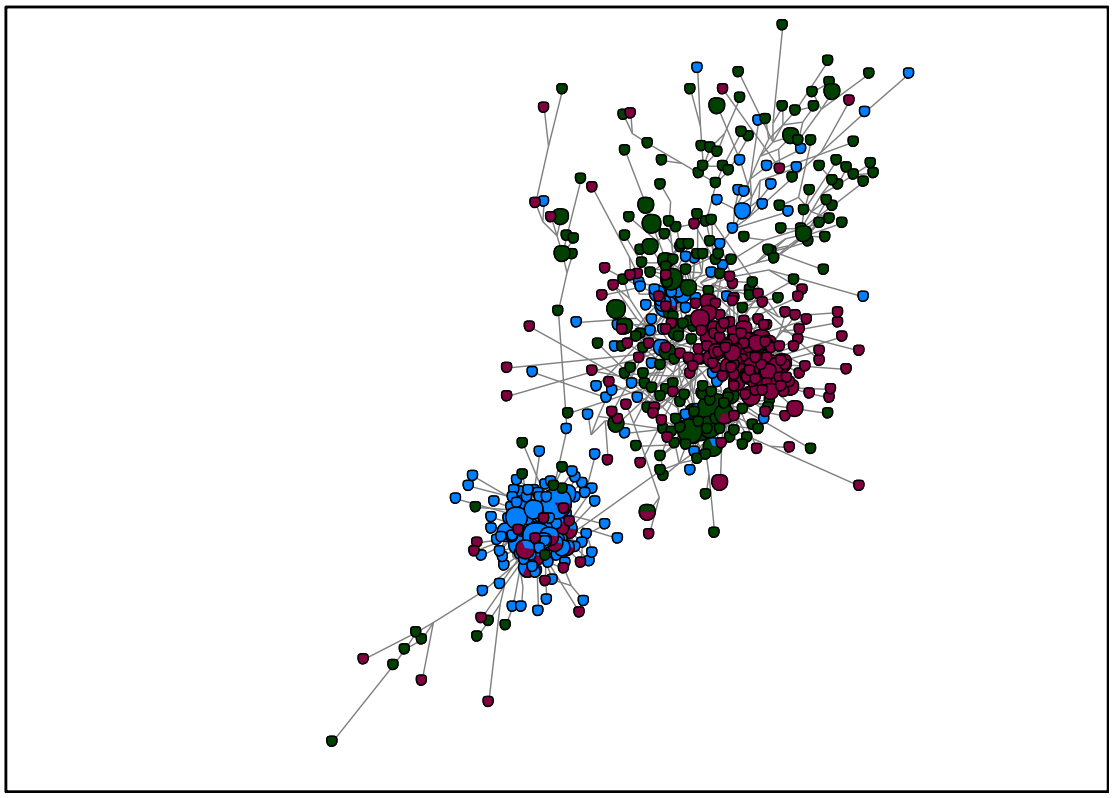


Figure 3.16 Torso of British population using weighted marker values
Caucasian haplotypes are shown in blue, Afro-Caribbean in red and South Asian in green. More importance was given to changes in the slowly mutating markers DYS392, DYS393 and DYS438 and less importance to changes in the more rapidly mutating marker DYS439.

3.5 Y-STR Population Determination

Both the striking differences at some individual loci, and the clustering of samples in Figure 3.16, suggests that it should be possible to design an algorithm for determining population-of-origin of an individual from a Y-STR haplotype. No system using Y-STR haplotypes will be able to provide 100% accuracy due to admixture throughout history (as shown by most, if not all, of the 10 shared haplotypes across populations) meaning that some individuals within a specific population will actually possess a Y chromosome originating from a different population. This is an inherent problem with using haploid markers, since assumptions on a person's origins are being made on the basis of the genetics of only one of many ancestors (down the male line: the father's, father's, father etc.) and this one ancestor may not be representative of the overarching genetic history of the individual. The quantified differentiation between the 3 population groups however indicates that it should be possible to produce an

efficient classification system with a known error rate even if total accuracy can never be attained.

A data mining approach, as implemented in the statistica software, was used to search the Y-STR data for consistent patterns between populations in order to develop a classification tree to predict sample group origin. The full 11-marker haplotypes of two hundred individuals from each of the 3 British population groups were used as a training set to develop the classification tree. Selection of the best approach was based on the lowest misclassification rate and tree simplicity. Validation of the classification system was then achieved by applying the algorithm to a new subset of data (50 individuals of each population group). For the developed model to be considered valid, the misclassification rate achieved with this additional data needed to be equal or less to that produced originally with the training set haplotypes. The classification validation was successful and the tree achieved can be seen in Figure 3.17.

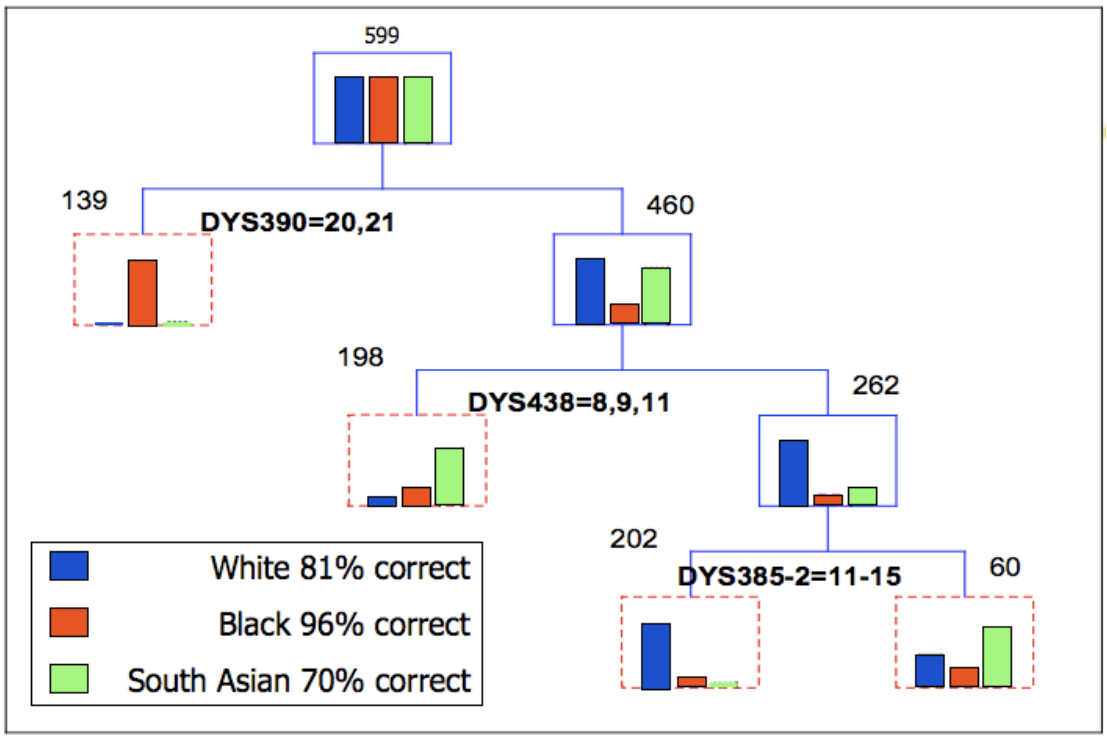


Figure 3.17 Y-STR population-of-origin classification system
Classification tree produced from the initial training sets of 200 individuals from each of the 3 British populations.

Figure 3.10 demonstrates that alleles 20 and 21 of locus DYS390 are nearly exclusively confined to the Afro-Caribbean population, and the algorithm uses this as the first classification step. Figure 3.17 displays the results from the original training set, however if all 750 samples are used then this 1st step classifies 178 samples as Afro-Caribbean, of which 2 are actually Caucasian and 8 are South Asian. Interestingly, 6 of these South Asian samples posses an 8 allele at locus DYS438: none of the 168 correctly allocated Afro-Caribbean samples have a DYS438 allele lower than a 10 and the overwhelming majority have an 11. The tree was designed to be simple, avoiding extra steps that were hard to validate and only separated out a few samples, however it would appear that these 6 South Asian samples can genuinely be identified by these changes in the slowly mutating DYS438 marker and are just included in this step due to an independently occurring mutation to allele 21 in DYS390.

The second step of the algorithm actually uses alleles 8 and 9 (and 11) of DYS438 to extract South Asian samples from those remaining haplotypes. If including all remaining samples from the 750, then this would classify 244 samples as South Asian, of which 22 are actually Caucasian and 37 Afro-Caribbean, giving this step a 76% success rate.

Step 3 uses the largest DYS385 allele to separate out Caucasian samples. Since individual locus typing of DYS385 is not performed (see Figure 1.12 for marker structure), it is not possible to know which allele is from the DYS385a locus and which from the DYS385b locus, so this step is purely looking at the larger of the 2 alleles present. A total of 251 samples are classified as Caucasian in this step, of which 28 are actually Afro-Caribbean and 16 are South Asian, giving the step a classification success of 83%.

There remain 77 samples (if looking at all 750 individuals) that the classification algorithm suggests assigning to the South Asian population group. Of these, 22 are Caucasian, 14 Afro-Caribbean and 41 South Asian – which would give a success rate for this step of only 53%.

When examining the samples classified in steps 2, 3 and 4 there would appear to be no major haplotypic differences between the correctly classified and misclassified samples. This would suggest that the wrongly assigned samples represent either a subsection of their population that has independently evolved to have similar Y-STR haplotype characteristics as the populations they are being classified into, or genuine admixture has occurred between the two populations at some time in history. One way to explore these two competing theories would be to examine slowly mutating Y-SNPs in these individuals to see whether similar Y-STR haplotypes are occurring on different Y-SNP backgrounds in the individuals from different populations.

Table 3.7 Likelihood ratios achieved for the predicted classification

Described / Predicted	Caucasian	Afro-Caribbean	South Asian
Caucasian		7 x	13 x
Afro-Caribbean	84 x		21 x
South Asian	5 x	4 x	

Numbers represent the likelihood ratios for how good the different classifications are (e.g. a sample classified as Caucasian is 7 times more likely to genuinely be Caucasian rather than Afro-Caribbean). This assumes prior odds of 1 (i.e. the unknown sample is equally likely to have come from all 3 groups).

Table 3.7 lists the likelihood ratios achieved if this classification system is followed. If the final 77 remaining samples are listed as unclassified rather than classified as South Asian then the likelihood ratios for the South Asian classified samples rise from 5 times more likely to be South Asian than Caucasian if classified as South Asian to 8 times, and from 4x more likely to be South Asian than Afro-Caribbean if classified as South Asian to 5x. This alteration does mean that only 90% of samples are now classified; however it does have the advantage of increasing the classification success rate. If the extra DYS438 step is added to purify the classified Afro-Caribbean sample of the 6 intruding South Asian samples, then the LR for Afro-Caribbean

classified samples jumps from being 21x more likely to be Afro-Caribbean than South Asian to 84x more likely.

These likelihood ratios are very useful for highlighting the success of the predictive steps, however the ratios are only applicable if the prior probability is 1: i.e. there is an equal chance that the unknown DNA sample could have come from any of the 3 populations groups. If it is suspected that the prior probability is not 1, then it is also worth taking note of the error rates for the different populations: 18% of Caucasian samples are wrongly classified, 33% of Afro-Caribbean samples, and 10% of South Asian samples. These rates fall to 27% for the Afro-Caribbean population and less than 10% for the Caucasian and South Asian data sets if the samples assigned in the last step of the algorithm are instead listed as non-classified.

Interest in determining the likely population of origin from a DNA sample is encountered in many fields, including adoption cases with young babies of unknown parentage, personal genealogy investigations and judicial matters. In criminal investigations this information is primarily desired by the police to direct investigations, for example to allocate resources in such a way that the investigation of a long list of potential suspect can to be prioritized. In this case, these likelihood ratios can prove very powerful, even with the classifications of lower confidence, in a similar way to forensic offender profiling.

4 Mitochondrial DNA Results and Discussion

The production and analysis of the mitochondrial data for this study was at times technically challenging and time consuming; from pyrosequencing primer design and assay optimisation, to manual haplogroup assignment and the generation and analysis of a minimum of 92 separate sequences for each full mitochondrial genome. In total the mitochondrial DNA of 537 individuals was sequenced in the course of this research.

Sequence changes with respect to the revised Cambridge reference sequence, and precise haplogroup affiliation, are listed below in Table 4.1. Some samples required sequence coverage of only the two hypervariable regions (HVI and HVII), while most samples were sequenced for the entire control region (approximately 1,100 bases). It was necessary to analyse extra SNPs within the mitochondrial coding region in order to obtain haplogroup assignments for some samples, and this is indicated in Table 4.2. Partial or complete sequencing of the coding region was also required for a select set of samples, and this is similarly denoted in Table 4.2. From here on, sequence data is mainly presented in phylogenetic networks in order to more clearly visualise the data.

A phylogenetic network is a way of showing how different genetic sequences relate to each other, for example if there are four sequences, which ones are most similar and which ones are most different. The Network software looks at the specific changes between these sequences and connects them together in a network. The most parsimonious network is that which requires the least number of *de novo* mutations to have occurred in order to connect all the sequences together. The more sequences there are, the easier it is to deduce a realistic phylogenetic tree by ordering the sequence changes in an evolutionary context (e.g. how are the different subspecies of grey wolf related to each other, and when did the grey wolf diverge with the coyote). This is especially easy to achieve with the non-recombining portion of the Y chromosome or with mitochondrial DNA, since the lack of recombination means that all mutations must occur in a sequential manner, and once a change has occurred and a branch point has been created in this phylogenetic tree, new mutations will occur independently on these two branches.

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Diagnostic changes from R*
Control Region
Full Coding & Control Region
Partial coding region

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)												HVII																
1	Jamaican	L0a1a2	129	148	168	172	187	188G	189	223	230	311	320			64	93	185	189	200	236	247	263	315.1C	523 del AC							
2	Jamaican	L0a1a2	129	148	168	172	187	188G	189	223	230	311	320	362	456	64	93	185	189	200	236	247	263	309.1C	315.1C	523 del AC						
3	Jamaican	L1b1*	126	187	189	223	264	270	278	293	311	519				73	152	182	185T	189	195	247	263	315.1C	357	523 del AC						
4	Jamaican	L1c2	129	187	189	223	265C	278	286G	294	311	360	519	527		73	151	152	182	186A	189C	195	198	247	263	264	297	315.1C	316			
5	Jamaican	L1c3	129	187	189	223	278	311	360							73	151	152	182	186A	189C	247	257	263	291T	297	315.1C	316	523 del AC			
6	Jamaican	L1c3a*	129	145	189	193.1C	223	278	294	311	360	519				73	151	152	182	186A	189C	247	263	315.1C	316	523 del AC						
7	Jamaican	L2a1	223	278	294	309	390	519								73	146	152	195	263	309.1C	315.1C										
8	Jamaican	L2a1*	223	278	294	309	390	519								73	143	146	152	195	239	263	309.1C	315.1C	523 del AC							
9	Jamaican	L2a1*	223	278	294	309	390									73	143	146	152	195	200	263	309.1C	315.1C	523 del AC							
10	Jamaican	L2a1*	86	189	223 C/T	278	294	309	390							73	146	152	195	263	315.1C											
11	Jamaican	L2a1*	183C	189	223	278	309	390								73	146	152	195	263	315.1C	534										
12	Jamaican	L2a1*	183 C/T	189	193.1C	223	278	294	309	390	519					73	146	152	195	263	309.1C	315.1C										
13	Jamaican	L2a1a2	223	278	286	294	309	390	519							73	146	152	195	263	309.1C	315.1C										
14	Jamaican	L2a1a2	93	192	223	278	286	294	309	390	519					73	146	152	195	263	315.1C											
15	Jamaican	L2a1c2a	193	213	223	239	278	294	309	390						73	143	146	152	195	263	309.1C	315.1C	513								
16	Jamaican	L2b1a	114A	129	213	223	278	355	362	390						73	150	152	182	195	198	204	257	263	315.1C	418	523 del AC					
17	Jamaican	L2b1a	114A	213	223	278	355	362	390							73	150	152	182	195	198	204	263	315.1C	523 del AC	573.1 Ins CCCC						
18	Jamaican	L2c*	223	278	390											73	93	150	152	182	186A	195	198	263	309.1C	309.2C	315.1C	325	523 del AC			
19	Jamaican	L2c3	167	223	235T	278	390									73	93	146	150	152	182	195	198	263	309.1C	315.1C	325	513	523 del AC			
20	Jamaican	L3b*	145	223	278	362	519									73	152	263	315.1C													
21	Jamaican	L3b*	223	278	291	362	519									73	263	309.1C	315.1C	523 del AC												
22	Jamaican	L3b*	124	223	278	362	519									73	152	263	309.1C	315.1C	523 del AC											
23	Jamaican	L3b1*	124	223	278	362	519									73	152	263	309.1C	315.1C	480	523 del AC										
24	Jamaican	L3b1*	124	150 C/T	223	278	291	362	519							73	152	263	309.1C	315.1C	523 del AC											
25	Jamaican	L3d*	124	189	223	362										71 het del G	73	152	263	315.1C	523 del AC											
26	Jamaican	L3d*	124	223												73	150	152	263	309.1C	315.1C	523 del AC										
27	Jamaican	L3d1a	124	223	319	519										73	152	195	263	315.1C												
28	Jamaican	L3d4	86	124	223	311	348									73	152	189	195	263	315.1C	523 del AC										
29	Jamaican	L3e1*	179	223	327	519										73	150	189	200	263	309.1C	315.1C										
30	Jamaican	L3e1*	179	223	327	519										73	150	152	189	200	204	263	315.1C									
31	Jamaican	L3e2a1	209	223	320	519										73	150	195	198	263	315.1C											
32	Jamaican	L3e2a1b	223	320	399	519										73	150	195	263	315.1C												
33	Jamaican	L3e2a1b1	86	223	320	399	519									73	150	195	198	263	315.1C	499										
34	Jamaican	L3e2b*	172	183C	186.1C	187	189	223	320	519						73	150	195	263	315.1C												
35	Jamaican	L3e2b*	172	189	223	320	519									73	150	195	263	315.1C												
36	Jamaican	L3e3	93	223	265T	519										73	150	195	263	315.1C	523 del AC											
37	Jamaican	L3e3	93	148	223	265T	519									73	150	195	263	315.1C	523 del AC											
38	Jamaican	L3e3	93	148	223	265T	519									73	150	195	263	315.1C	523 del AC											
39	Jamaican	L3e3	93	223	265T	519										73	150	195	263	309.1C	315.1C	523 del AC										
40	Jamaican	L3f1b*	209	223	287	292	311	519								73	150	189	200	263	309.1C	315.1C										
41	Jamaican	L3f1b*	209	223	263	311	368	519								73	150	189	200	263	309.1C	315.1C	523 del AC									
42	Jamaican	L3f1b1	129	209	223	292	295	311	519							73	189	200	263	315.1C	523 del AC											
43	Jamaican	L3f1b1	129	209	223	292	294	295	311	519						73	189	200	263	309.1C	315.1C											
44	Jamaican	L3f1b1	129	209	223	292	295	311	519							73	150	189	200	207	263	309.1C	315.1C									
45	Barbadian	H1c13	519													263	315.1C	477														
46	Barbadian	HV0*	298	301												72	263	309.1C	315.1C													
47	Barbadian	L0a1a	126	129	148	168	172	187	188G	189	223	230	311	320	519	93	185	189	200	236	247	263	309.1C	315.1C	523 del AC							
48	Barbadian	L1b	126	187	189	223	264	270	278	293	311	519				73	152	182	185T	195	247	263	315.1	357	523 del AC							
49	Barbadian	L1b	93	126	145	187	189	223	264	270	278	293	311	399	519	73	152	182	185T	195	247	263	315.1C	357	523 del AC							
50	Barbadian	L1b	93	126	145	187	189	223	264	270	278	293	311	399	519	73	152	182	185T	195	247	263	315.1C	357	523 del AC							
51	Barbadian	L1b	126	187	189	223	264	270	278	311	519					73	152	182	185T	195	247	263	315.1C	357	523 del AC							
52	Barbadian	L1b*	93	111	126	187	189	223	239	270	278	293	311	519		73	146	152	182	185T	189	247	263	315.1C	357	523 del AC						
91	Barbadian	L1b*	37	126	172	187	189	223	264	270	278	293	301	311	519	73	152	182	185C	189	195	247	263	315.1C	357	523 del AC						
53	Barbadian	L1b1a*	126	187	189	223	264	270	278	311	519					73	152	182	185T	189	195	247	263	315.1C	357	523 del AC						
54	Barbadian	L1b1a*	126	187	189	223	264	270	278	293	311	519				73	152	185T	189	195	247	263	315.1C	357	523 del AC							
55	Barbadian	L1b1a*	126	187	189	223	264	266	270	278	293	311	519			73	152	185T	189	195	207	247	263	315.1C	357	523 del AC						
56	Barbadian	L1c*	114G	129	187	189	223	261	278	311	360	519				73	151	152	182	186A	189C	247	263	291T	297	309.1C	315.1C	316	523 del AC			
57	Barbadian	L1c2	93	129	187	189	223	265C	278	286G	294	311	360	519	527	73	151	152	182	186A	189C	195	198	247	263	297	315.1C	316	523 del AC			
58	Barbadian	L1c2	51	129	187	189	223	265C	278	286G	288	294	311	360	519 527	73	151	152	182	186A	189C	195	198	247	263	297	315.1C	316				
59	Barbadian	L1c2b	129	187	189	223	230	265C	278	286A	294	311	519	527		73	150	151	152	182	186A	189C	195	198	247	263	297	309.1C	315.1C	316		
60	Barbadian	L2a1*	189	192	223	278	294	362	390							73	143	146	152	195	263	309.1C	315.1C									
61	Barbadian	L2a1*	189	223	278	294	309	390								73	146	152	195	263	315.1C											
62	Barbadian	L2a1*	129	189	223	278	294	309	390							73	143	146	152	263	315.1C	523 del AC										
63	Barbadian	L2a1*	183C	189	223	278	294	309	362	390						73	146	152	195	263	309.1C	315.1C										
64	Barbadian	L2a1*	189	192	223	278	294	309	390	519						73	146	152	195	263	315.1C											
67	Barbadian	L2a1a2	92	223	256	278	286	294	309	390	519					73	146	195	263	315.1C												

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)													HVII														
68	Barbadian	L2a1a2	92	223	256	278	286	294	309	390	519					73	146	195	263	315.1C											
69	Barbadian	L2a1a2	92	223	278	286	294	309	390	519						73	146	152	195	263	309.1C	315.1C									
95	Barbadian	L2a1a2	92	223	256	278	286	294	309	390	519					73	146	195	263	315.1C											
70	Barbadian	L2a1i	189	192	203	223	278	294	362	390						73	143	146	152	195	263	315.1C									
71	Barbadian	L2a1i	86	129	189	193.1C	223	278	294	362	390					73	143	146	152	195	263	309.1C	309.2C	315.1C							
72	Barbadian	L2a1i	189	191.1C	192	223	278	294	362	365	390					73	143	146	152	195	204	263	315.1C								
73	Barbadian	L2a1i	189	191.1C	192	223	278	294	362	390						73	143	146	152	195	263	315.1C	524 ins AC								
74	Barbadian	L2b	51	114A	129	213	223	278	390							73	146	150	152	182	185C	195	198	204	263	309.1C	315.1C				
75	Barbadian	L2c	223	278	390											73	93	95C	146	150	152	182	185C	195	198	263	309.1C	315.1C	325	523 del AC	
76	Barbadian	L2c	189	223	278	390										64	73	93	146	150	182	195	198	263	309.1C	315.1C	325	523 del AC			
77	Barbadian	L3b*	145	223	278	362	519									73	152	263	315.1C	523 del AC											
78	Barbadian	L3b1*	124	223	278	362	519									73	152	263	309.1C	315.1C	523 del AC										
79	Barbadian	L3d1c	124	166	223											73	152	263	315.1C	523 del AC											
80	Barbadian	L3d1c	124	166	223											73	152	263	315.1C	523 del AC											
81	Barbadian	L3e1a	185	223	327	519										73	150	189	200	263	309.1C	315.1C									
82	Barbadian	L3e2a1b1	145	223	320	399	519									73	150	195	198	263	315.1C	499									
83	Barbadian	L3e2b*	172	183C	189	223	320	519								73	150	152	195	263	309.1C	315.1C									
85	Barbadian	L3f1b1	129	209	223	292	295	311	519							73	185	189	200	263	309.1C	315.1C									
99	Barbadian	L4b1	179	183C	189	223	239	311	320	362	519					73	150	199	204	263	315.1C	507	513								
86	Barbadian	L4b2	93G	223	287A	293T	301	311	355	362	399					73	146	150	152	195	200	207	244	263	315.1C	513					
87	Barbadian	L4b2	93G	223	287A	293T	301	311	355	362	399					73	146	150	152	195	200	207	244	263	315.1C	513					
88	Barbadian	U6a3b	172	183C	189	219	278	519								73	146	185	188	263	291.1A	309.1C	315.1C	523 del AC							
89	Grenada	L0a1a2	129	148	168	172	187	188G	189	223	230	291	311	320		64	93	185	189	200	236	247	263	315.1C	523 del AC						
90	St Vincent	L1b*	126	187	189	223	264	270	278	293	311	519				73	152	182	185T	189	195	247	263	309.1C	315.1C	357	466	523 del AC			
92	St Lucia	L1c*	86	129	187	189	223	241	278	291	294	311	519			73	151	152	182	186A	189C	195	198	247	263	297	315.1C	316	523 del AC		
93	Grenada	L2a1	223	278	294	309	390	519								73	146	152	195	217	263	309.1C	315.1C								
94	Grenada	L2a1*	183C	189	223	278	294	309	390							73	146	152	195	263	309.1C	315.1C									
65	St Lucia	L2a1*	189	223	278	294	309	390								73	146	152	195	263	309.1C	315.1C									
66	St Lucia	L2a1a1	51	223	278	294	297A	309	368	390	519					73	146	152	195	263	309.1C	315.1C	524 ins ACAC								
96	Trinidad	L2b*	114A	129	213	223	278	390								73	146	150	152	182	183	186	195	198	204	263	309.1C	315.1C			
97	Trinidad	L2b1a	114A	129	145	213	223	256	278	355	362	390				73	150	152	182	195	198	204	263	315.1C	418	523 del AC					
98	St Lucia	L2b1a	114A	129	213	223	261	278	355	362	390					73	150	152	182	195	198	204	228	263	315.1C	418	523 del AC				
84	Guyana	L3e3	223	265T	519											73	150	195	263	309.1C	315.1C	523 del AC									
100	Trinidad	M7c3c	223	295	362	519										73	146	199	263	309.1C	315.1C	489	523 del AC								
101	Afro-Caribbean	HV1b1	67	158	274											263	315.1C														
102	Afro-Caribbean	K1a	224	311	519											73	263	309.1C	315.1C	497											
103	Afro-Caribbean	L0a1*	92	129	148	168	172	187	188G	189	223	230	311	320		93	152	185	189	200	236	247	263	315.1C	523 del AC						
104	Afro-Caribbean	L0a1b1	129	148	168	172	187	188G	189	223	230	278	293	311	320	519	93	95C	185	189	236	247	263	309.1C	315.1C	523 del AC					
105	Afro-Caribbean	L0a2	148	172	187	188G	189	223	230	311	320						64	93	152	236	263	309.1C	315.1C								
106	Afro-Caribbean	L1b	126	187	189	223	264	270	278	311	519						73	152	182	185T	195	247	263	315.1C	357	523 del AC					
107	Afro-Caribbean	L1b	126	187	189	223	264	270	278	311	519						73	151	152	182	185T	195	247	263	315.1C	357	523 del AC				
108	Afro-Caribbean	L1b*	126	166 del A	172	183C	223	264	270	278	293	311	519				73	146	152	182	185T	195	247	263	315.1C	357	523 del AC				
109	Afro-Caribbean	L1b*	126	187	189	223	264	278	293	311	519						73	152	182	185T	247	263	315.1C	357	523 del AC						
110	Afro-Caribbean	L1b*	126	168	187	189	223	264	270	278	311						73	152	182	185T	195	247	263	309.1C	315.1C	357					
111	Afro-Caribbean	L1b*	126	187	189	223	264	270	278	311							73	152	182	185T	195	247	263	315.1C	357						
112	Afro-Caribbean	L1b*	126	187	189	223	264	270	278	293	311						73	152	182	185T	195	247	263	315.1C	357						
113	Afro-Caribbean	L1b*	126	187	189	223	264	270	278	293	311						73	152	182	185T	195	247	263	309.1C	315.1C	357					
114	Afro-Caribbean	L1b*	126	187	189	223	264	270	278	293	311						73	152	182	185T	195	247	263	315.1C	357						
115	Afro-Caribbean	L1b1*	126	187	189	223	264	270	278	293	311	519					73	152	182	185T	189	195	247	263	309.1C	315.1C	357	522 del CA			
116	Afro-Caribbean	L1b1a	126	187	189	223	264	270	278	293	311	519					73	152	182	185T	189	195	247	263	309.1C	315.1C	357	523 del AC			
117	Afro-Caribbean	L1b1a	126	165T	187	189	223	264	270	278	293	311	519				73	152	182	185T	189	195	247	263	315.1C	357	523 del AC				
118	Afro-Caribbean	L1b1a	126	187	189	223	264	270	278	293	311						73	152	185T	189	195	247	263	315.1C	357						
119	Afro-Caribbean	L1b1a	126	187	189	223	264	270	278	311							73	152	182	185T	189	195	247	263	315.1C	357					
120	Afro-Caribbean	L1b1a	126	187	189	223	264	270	278	293	311						73	152	182	185T	189	195	247	263	315.1C	357					
121	Afro-Caribbean	L1b1a	126	187	189	223	264	270	278	311							73	152	182	185T	189	195	247	263	309.1C	315.1C	357				
122	Afro-Caribbean	L1b2	111	126	187	189	*223*	239	270	278	293	311	519				73	146	152	182	185T	189	247	263	315.1C	357	523 delAC				
123	Afro-Caribbean	L1c*	129	187	189	223	261	278	311	360	519						73	151	152	186A	189C	247	263	291T	297	309.1C	315.1C	316	523 del AC		
124	Afro-Caribbean	L1c1*	86	129	187	189	223	278	293	294	311	360					73	151	152	182	186A	189C	195	198	247	263	297				
125	Afro-Caribbean	L1c1a1	129	162	172	173	188A	189	223	256	278	293	294	311	360	368	519	73	151	152	182	186A	189	195	198	247	263	297	315.1C	316	523-4del ac
126	Afro-Caribbean	L1c2b	129	187	189	223	265C	278	286G	294	311	360	527				73	151	152	182	186A	189C	195	198	247	263	297	315.1C	316	471	523 del AC
127	Afro-Caribbean	L1c2b	129	187	189	214	223	265C	278	286A	291	294	311	360			73	151	152	182	186A	189C	195	198	247	263	297	315.1C	316		
128	Afro-Caribbean	L1c3a	93	129	183C	189	193.1	215	278	294	311	360					73	152	182	186A	189C	198	247	263	309.1C	315.1C	316				
129	Afro-Caribbean	L1c3a	129	182C	183C	189	193.1	193.2C	215	223	278	294	311	360			73	151	152	182	186A	189C	195	247	263	315.1C	316				
130	Afro-Caribbean	L1c3a1b	182C	183C	189	193.1C	215	223	231	278	294	311	355	360	519		73	151	152	182	186A	189C	247	263	315.1C	316	523 del AC				
131	Afro-Caribbean	L2a*	183C	189	193.1C	193.2C	223	278	291	294	390						73	143	146	152	195	263	315.1C								
132	Afro-Caribbean	L2a*	183C	189	193.1C	223	278	291	294	390							73	143	146	152	195	263	315.1C								
133	Afro-Caribbean	L2a1	223	278	294	309	390										73	146	152	195	263	315.1C									
134	Afro-Caribbean	L2a1	223	278	294	309	390										73	146	152	195	263	309.1C	315.1C								
135	Afro-Caribbean	L2a1	223	278	294	309	390										73	146	152	195	263	309.1C	315.1C								
136	Afro-Caribbean	L2a1*	223	278	294	311	355	390	519								73	143	146	152	195	263	315.1C								
137	Afro-Caribbean	L2a1*	223	278	294	309	362	390	519								73	146	152	195	196	263	309.1C	315.1C							

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)													HVII													
138	Afro-Caribbean	L2a1*	189	192	223	278	294	309	390	519						73	146	152	195	263	309.1C	315.1C								
139	Afro-Caribbean	L2a1*	129	183C	189	192	223	278	294	309	390					73	143	146	152	195	263	309.1C	315.1C							
140	Afro-Caribbean	L2a1*	189	192	223	278	294	390								73	146	152	195	263	315.1C									
141	Afro-Caribbean	L2a1*	223	256	278	294	309	390								73	143	146	152	195	263	315.1C	339							
142	Afro-Caribbean	L2a1*	129	189	223	278	294	309	390							73	143	146	152	263	315.1C									
143	Afro-Caribbean	L2a1*	86	223	278	294	309	390								73	146	152	195	263	309.1C	315.1C								
144	Afro-Caribbean	L2a1*	189	223	278	294	309	390								73	143	146	152	195	263	315.1C								
145	Afro-Caribbean	L2a1*	129	189	192	223	278	294	309	390						73	143	146	152	195	263	315.1C								
146	Afro-Caribbean	L2a1*	182C	183C	189	192	223	278	290	294	309	390				73	146	152	195	263	315.1C									
147	Afro-Caribbean	L2a1*	93	189	193.1C	223	278	294	309	390						73	146	152	195	263	309.1C	315.1C								
148	Afro-Caribbean	L2a1*	189	223	278	292	294	309	356	390						73	143	146	152	195	263	315.1C								
149	Afro-Caribbean	L2a1*	93	185	189	191.1C	192	223	278	294	309	360	390			73	146	152	195	198	263	315.1C								
150	Afro-Caribbean	L2a1a1	223	278	294	309	368	390	519							73	146	152	195	198	263	292.1A	294.1T	309.1C	315.1C	523.1C	523.2A	523.3C	523.4A	
151	Afro-Caribbean	L2a1a1	223	278	294	309	368	390								73	146	152	195	198	263	309.1C	315.1C							
152	Afro-Caribbean	L2a1a2	223	278	286	294	309	390	519							73	146	152	195	263	309.1C	315.1C								
153	Afro-Caribbean	L2a1c1	86	223	278	294	309	390								73	143	146	152	195	198	263	315.1C							
154	Afro-Caribbean	L2a1c2a	93	193	213	223	239	278	294	309	390					73	143	146	152	195	263	315.1C								
155	Afro-Caribbean	L2a1c5	129	148	223	278	294	390								73	143	146	152	195	263	309.1C	315.1C							
156	Afro-Caribbean	L2a1l	189	223	278	294	309	390	519							73	143	146	152	195	263	315.1C	534							
157	Afro-Caribbean	L2b*	114A	129	213	223	278	390								73	146	150	152	182	183	186	195	198	204	263	309.1C	315.1C		
158	Afro-Caribbean	L2b1a	114A	129	213	223	278	355	362	368	390					73	150	182	195	198	263	315.1C	418	523-4del ac						
159	Afro-Caribbean	L2b1a	114A	129	213	223	278	355	362	390						73	152	182	195	198	204	263	315.1C							
160	Afro-Caribbean	L2b1b	114A	129	213	223	278	325	362	390						73	146	150	152	153	182	183	195	198	204	263	315.1C	385	418	523 del AC
161	Afro-Caribbean	L2b1b	114A	129	213	223	278	362	390							73	146	150	152	182	183	195	198	204	263	309.1C	315.1C	385		
162	Afro-Caribbean	L2c*	189	223	278	362	390									73	93	146	150	152	182	195	198	263	315.1C	325	455.1T	523 del AC		
163	Afro-Caribbean	L2c*	223	278	390	519										73	146	150	152	182	195	199	263	309.1C	315.1C	325				
164	Afro-Caribbean	L2c*	223	278	311	390	519									73	89	93	146	150	152	182	195	198	263	315.1C	325	523 del AC		
165	Afro-Caribbean	L2c*	223	278	390											73	93	146	150	152	182	195	198	263	315.1C	325				
166	Afro-Caribbean	L2c*	193	223	278	390										73	93	146	150	152	182	195	198	263	315.1C	325				
167	Afro-Caribbean	L2c*	223	278	390	399										73	93	146	150	152	182	195	198	263	315.1C	325				
168	Afro-Caribbean	L2c2	223	264	278	390										73	93	146	150	152	182	195	198	263	315.1C	325				
169	Afro-Caribbean	L2c2b	223	278	390											73	89	93	146	150	152	182	183	195	198	263	315.1	325		
170	Afro-Caribbean	L2e	111A	145	184	189	223	239	278	292	355	362	390	399	400	73	151	152	182	185	263	309.1C	315.1C							
171	Afro-Caribbean	L3b*	48	124	223	278	362									73	152	185	189	263	315.1C	523 del AC								
172	Afro-Caribbean	L3b*	124	223	270	278	311	362	519							73	263	315.1C	523 del AC											
173	Afro-Caribbean	L3b*	93	124	278	362	519									73	263	315.1C	523 del AC											
174	Afro-Caribbean	L3b*	124	223	278	519										73	263	315.1C	523 del AC											
175	Afro-Caribbean	L3b*	124	223	274	278	519									73	263	315.1C	393	523 del AC										
176	Afro-Caribbean	L3b*	124	223	224	278	362									73	195	263	315.1C											
177	Afro-Caribbean	L3b*	223	278	362											73	185	189	263	315.1C										
178	Afro-Caribbean	L3b*	124	223	278	362										73	195	263	309.1C	315.1C										
179	Afro-Caribbean	L3b*	124	223	278	362										73	263	315.1C												
180	Afro-Caribbean	L3b*	124	223	278	355	362									73	263	315.1C												
181	Afro-Caribbean	L3b*	124	223	278	362										73	263	315.1C												
182	Afro-Caribbean	L3d*	124	223	256	362										73	146	152	263	309.1C	315.1C	523 del AC								
183	Afro-Caribbean	L3d*	124	153	223	291										73	152	263	309.1C	315.1C	523 del AC									
184	Afro-Caribbean	L3d*	124	223	325	327										73	146	152	263	315.1C										
185	Afro-Caribbean	L3d1a	124	223	319											73	152	263	309.1C	315.1C	523 del AC									
186	Afro-Caribbean	L3d1b3	124	223												73	146	152	263	309.1C	315.1C									
187	Afro-Caribbean	L3d1b3	124	223												73	146	152	263	315.1C										
188	Afro-Caribbean	L3d1b3	124	223												73	146	152	263	309.1C	315.1C									
189	Afro-Caribbean	L3d1d	124	223	256	278	368	399								73	152	263	309.1C	315.1C	523 del AC									
190	Afro-Caribbean	L3d1d	124	223	256	368										73	152	263	315.1C											
191	Afro-Caribbean	L3d4	124	223	260											73	152	189	195	263	315.1C									
192	Afro-Caribbean	L3e1	223	311	327											73	150	189	200	263	309.1C	315.1C	523-4del ac							
193	Afro-Caribbean	L3e1*	223	327												73	150	189	195	200	263	309.1	315.1C							
194	Afro-Caribbean	L3e1*	223	327												73	150	152	189	200	263	309.1C	315.1C							
195	Afro-Caribbean	L3e1*	223	327												73	150	189	195	200	263	309.1C	315.1C							
196	Afro-Caribbean	L3e1a2	129	185	209	223	327									73	150	152	189	195	200	207	263	315.1C							
197	Afro-Caribbean	L3e2	223	320												73	150	195	263	315.1C										
198	Afro-Caribbean	L3e2	223	320												73	150	195	263	309.1C	315.1C									
199	Afro-Caribbean	L3e2*	92	223	258T	320	519									73	150	189	195	263	309.1C	315.1C								
200	Afro-Caribbean	L3e2a1	223	320												73	150	195	198	263	315.1C									
201	Afro-Caribbean	L3e2a1b1	223	301	320	399										73	150	195	198	263	315.1C									
202	Afro-Caribbean	L3e2a1b2	189	209	223	311	320	519								73	150	195	198	263	309.1C	315.1C								
203	Afro-Caribbean	L3e2a1b2	124	223	311	320										73	150	195	198	263	315.1C									
204	Afro-Caribbean	L3e2b	189	223	258T	320	519									73	150	189	195	263	309.1C	315.1C								
205	Afro-Caribbean	L3e2b	172	183C	189	223	230	311	320	519						73	150	195	263	309.1C	315.1C									
206	Afro-Caribbean	L3e2b	172	189	223	320										73	150	195	263	282	315.1C									
207	Afro-Caribbean	L3e2b	172	183C	189	223	320									73	150	195	263	309.1C	309.2C	315.1C								
208	Afro-Caribbean	L3e2b	172	189	223	311	320									73	150	195	263	309.1C	315.1C									
209	Afro-Caribbean	L3e2b*	126T/C	172	183C	189	213	223	320	519						73	150	152	195	263	309.1C	315.1C								
210	Afro-Caribbean	L3e2b*	172	183C	189	223	320	519								73	150	152	195	263	315.1C									

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)								HVII										
211	Afro-Caribbean	L3e2b*	172	183C	189	223	320				73	150	152	195	263	315.1C					
212	Afro-Caribbean	L3e2b*	172	183C	189	193.1C	223	320			73	150	152	195	263	315.1C					
213	Afro-Caribbean	L3e3	265T	519							73	150	195	263	309.1C	315.1C	523 del AC				
214	Afro-Caribbean	L3e3	93	223	265T						73	150	195	263	315.1C						
215	Afro-Caribbean	L3e3	93	148	223	265T					73	150	195	263	309.1C	315.1C					
216	Afro-Caribbean	L3e3	93	223	265T						73	150	195	263	315.1C						
217	Afro-Caribbean	L3e4	51	223	264	519					73	150	263	315.1C	523 del AC						
218	Afro-Caribbean	L3e4	51	223	264						73	150	263	309.1C	315.1C						
219	Afro-Caribbean	L3e4	51	223	264						73	146	150	263	309.1C	315.1C					
220	Afro-Caribbean	L3f1a	126	209	223	519					73	228	263	315.1C							
221	Afro-Caribbean	L3f1b*	189	193.1C	209	223	292	311	519		73	150	189	200	263	309.1C	315.1C				
222	Afro-Caribbean	L3f1b*	209	223	292	311	519				73	150	189	200	263	309.1C	315.1C				
223	Afro-Caribbean	L3f1b*	209	223	292	311					73	150	189	263	315.1C						
224	Afro-Caribbean	L3f1b1	129	209	223	292	295	311			73	152	189	263	315.1C						
225	Afro-Caribbean	L3f1b1	129	209	223	292	295	311			73	189	200	263	309.1C	315.1C	385				
226	Afro-Caribbean	L3f1b1	129	209	223	292	295	311			73	189	200	263	272	309.1C	315.1C				
227	Afro-Caribbean	L3f1b1	129	209	223	292	295	311			73	189	200	207	263	309.1C	315.1C				
228	Afro-Caribbean	L3f1b4a	209	223	311						73	150	189	200	263	309.1C	315.1C				
229	Afro-Caribbean	L3f1b4a	209	223	311						73	150	189	200	263	309.1C	315.1C				
230	Afro-Caribbean	L3k	223	355							73	150	152	235	263	315.1C	494				
231	Afro-Caribbean	L3k	223	355							73	150	152	235	263	315.1C					
232	Afro-Caribbean	M1a1	129	189	193.1CC	223	249	311	359	519	73	150	189	195	198	263	309.1C	309.2C	315.1C	489	523 del AC
233	Afro-Caribbean	M65a	223	289	519						73	146	263	315.1C	489	511					
234	Afro-Caribbean	M6a	223	231	356	362	519				73	263	315.1C	461	489	523 del AC					
235	Afro-Caribbean	U8a1	146	209	342						73	263	282	309.1C	315.1C						
236	Caucasian	A4b	39	188	189	223	290	319	356	362	73	152	235	263	309.1 C	315.1 C					
237	Caucasian	D4i	223	294	362						73	146	195	263	309.1C	315.1C	379	489			
238	Caucasian	H*	222	311	519						150	263	315.1C								
239	Caucasian	H*	93								200	263	315.1 C								
240	Caucasian	H*	519								151	152	198	263	309.1C	315.1C					
241	Caucasian	H1	Anderson								263	309.1 C	309.2 C	315.1 C							
242	Caucasian	H1*	519								152	199	263	309.1C	315.1C	471	524 ins ACAC				
243	Caucasian	H1*	519								146	263	309.1C	315.1C							
244	Caucasian	H1*	80	189	209	356					263	309.1C	315.1C								
245	Caucasian	H1*	93	519							146	263	315.1C								
246	Caucasian	H1*	129	519							93	263	309.1C	315.1C							
247	Caucasian	H1*	183C	189	519						263	309.1C	315.1C								
248	Caucasian	H1*	519								263	309.1C	309.2C	315.1C							
249	Caucasian	H1*	239								263	315.1 C									
250	Caucasian	H1*	294	304							263	315.1 C									
251	Caucasian	H1*	Anderson								195	257	263	309.1C	315.1 C						
252	Caucasian	H1*	Anderson								152	263	315.1 C								
253	Caucasian	H1*	Anderson								263	315.1 C									
254	Caucasian	H1*	Anderson								146	263	315.1 C								
255	Caucasian	H1*	51								152	263	315.1 C								
256	Caucasian	H10e	93	148	221	519					263	309.1C	315.1C								
257	Caucasian	H11a	293	311							195	263	315.1 C								
258	Caucasian	H11a1	278	293	311						143	195	263	315.1 C							
259	Caucasian	H15a1	311								55	57G	263	309.1C	315.1C						
260	Caucasian	H16c	519								152	263	315.1C								
261	Caucasian	H17c	92	239	519						73	263	315.1 C								
262	Caucasian	H1a1	162	209							73	263	315.1 C								
263	Caucasian	H1b	172	189	356	362	519				152	263	315.1C	523 del AC							
264	Caucasian	H1ba	270	519							257 A/G	263	315.1C								
265	Caucasian	H1c1	263	519							152	263	315.1C	477							
266	Caucasian	H24a	168	293	519						263	315.1 C									
267	Caucasian	H27	129	316	519						263	315.1C									
268	Caucasian	H2a*	Anderson								263	309.1	315.1	523 del AC							
269	Caucasian	H2a1	172	354							263	315.1 C									
270	Caucasian	H2a1d	354								189 A/G	263	309.1C	315.1C							
271	Caucasian	H2a1f	193	354							127	263	309.1 C	315.1 C							
272	Caucasian	H3*	519								263	310delT									
273	Caucasian	H3*	519								263	310del									
274	Caucasian	H3*	289	519							152	263	309.1C	309.2C	315.1C	524 ins AC					
275	Caucasian	H3*	Anderson								263	315.1 C									
276	Caucasian	H3*	311								73	263	315.1 C								
277	Caucasian	H3as	93	519							263	315.1 C									
278	Caucasian	H4	Anderson								73	263	292 C/T	309.1C	315.1C	523 del AC					
279	Caucasian	H5	304								152	263	309.1C	315.1C	456	523 del AC					
280	Caucasian	H5	304								263	315.1C	338	456							
281	Caucasian	H51	519								263	309.1 C	315.1 C								
282	Caucasian	H6a*	193	219	362	482					204	239	263	309.1C	315.1C						
283	Caucasian	H6a*	288	362	482	519					239	263	309.1C	309.2C	315.1C						
284	Caucasian	H6a*	362								239	263	309.1 C	315.1 C							
285	Caucasian	H6a1b2	93T/C	362	482						239	263	309.1C	315.1C							
286	Caucasian	H7*	Anderson								152	263	315.1 C								

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)										HVII										
287	Caucasian	HV0*	298	301									72	249delA	263	309.1C	315.1C						
288	Caucasian	HV0*	298										72	152	263	309.1 C	309.2 C	315.1 C					
289	Caucasian	HV16	311										263	309.1 C	315.1 C								
290	Caucasian	HV5	Anderson										263	315.1 C									
291	Caucasian	I*	129	223	362	391							73	152	199	204	207	250	263	309.1 C	309.2 C	315.1 C	
292	Caucasian	I1a1	129	172	223	311	391						73	199	203	204	250	263	315.1 C				
293	Caucasian	I1a1	129	172	223	311	391						73	199	203	204	250	263	309.1 C	315.1 C			
294	Caucasian	J1b1a	69	126	145	172	222	261					10	73	146	242	263	295	315.1C	462	489		
295	Caucasian	J1b1a1	69	126	145	172	215	222	261				73	242	263	295	315.1 C						
296	Caucasian	J1b1a1a	69	126	145	172	192	222	261				73	242	263	295	315.1C	462	489				
297	Caucasian	J1c1	69	126	193								73	185	228	263	295	315.1C	462	482	489		
298	Caucasian	J1c2	69	126	258 C								73	150	185	188	228	263	295	315.1 C			
299	Caucasian	J1c2	69	126	519								73	152	185	188	228	263	295	309.1C	315.1C	462	489
300	Caucasian	K1a	224	311	519								73	263	315.1 C	497							
301	Caucasian	K1a*	93	224	311	519							73	263	315.1 C	497							
302	Caucasian	K1a*	189	224	311	519							73	263	309.1C	315.1C	497						
303	Caucasian	K1a1	224	311	519								73	114	152	263	315.1 C	497					
304	Caucasian	K1a4a1a2b	224	245	311	519							73	146	263	309.1 C	309.2 C	315.1 C	497	524.1 Ins AC			
305	Caucasian	K1a4a1a2b	224	245	311	519							73	146	263	315.1C	497						
306	Caucasian	K1b1a	153	224	311	319	463	519					73	152	195	263	315.1 C	524 ins AC					
307	Caucasian	K1c	224	311	519								73	146	152	263	315.1 C	498del					
308	Caucasian	K1c	129	224	311	519							73	146	152	263	315.1C	498del					
309	Caucasian	L1b	126	187	189	223	264	270	278	293	311	519	73	152	182	185 T	195	247	263	315.1 C	357	523 del AC	
310	Caucasian	M2c	223	258 del A	274								73	199	263	309.1C	315.1C	447G	489	523 del AC			
311	Caucasian	N1a1a3	129	147A	154	172	223	248	320	355	519		73	152	199	204	234	263	309.1C	315.1C	573.1CCCCC		
312	Caucasian	T1a	126	163	186	189	294						73	152	195	263	309.1 C	315.1 C					
313	Caucasian	T2*	126	294	296	519							44.1C	73	263	315.1C							
314	Caucasian	T2*	126	219	294	296							73	263	309.1 C	315.1 C							
315	Caucasian	T2b	126	294	296	304							73	263	315.1C								
316	Caucasian	T2b	126	294	296	304							73	309.1 C	315.1 C								
317	Caucasian	T2c1b	126	147	292	294	519						73	146	152	263	279	315.1C					
318	Caucasian	U5a1*	183 del	256	270	362	399						73	263	315.1 C								
319	Caucasian	U5a2	192	256	270	526							73	263	315.1 C								
320	Caucasian	U5a2a	114 A	256	270	294	526						73	263	309.1 C	315.1 C	524.1 Ins AC						
321	Caucasian	U5a2a	114 A	192	256	270	294	350	526				73	152	263	309.1 C	315.1 C	372					
322	Caucasian	U5b*	192	270									73	150	263	315.1 C							
323	Caucasian	U5b2a1b	189	193.1C	325								73	150	152	263	309.1C	315.1C					
324	Caucasian	X2c1a	108	182C	183C	189	223	255	278	519			73	153	195	225	227	263	309.1C	309.2C	315.1C		
325	Irish	H*	129	519									93	200	263	309.1C	315.1C						
326	Irish	H*	274	519									263	315.1 C									
327	Irish	H1*	519										204	263	315.1C								
328	Irish	H1*	111 A	519									263	309.1C	315.1C								
329	Irish	H1*	Anderson										152	263	315.1 C								
330	Irish	H1*	93										263	309.1 C	315.1 C								
331	Irish	H1*	311										263	315.1 C									
332	Irish	H1*	291										263	315.1 C									
333	Irish	H1*	129										152	263	309.1 C	315.1 C							
334	Irish	H1*	Anderson										263	315.1 C									
335	Irish	H1*	Anderson										195 A	253 T/C	263	315.1 C							
336	Irish	H1a1	162	209									73	263	315.1 C								
337	Irish	H1b	189	209	356	362							263	315.1 C									
338	Irish	H1b*	51	75	189	311	356	519					263	315.1C									
339	Irish	H1i2*	519										152	263	315.1 C								
340	Irish	H2a2a	235										263	309.1C	315.1C								
341	Irish	H2a2a1	Anderson										315.1 C										
342	Irish	H3*	519										152	263	309.1 C	309.2 C	315.1 C						
343	Irish	H3*	519										152	263	309.1C	315.1C							
344	Irish	H3*	129	256	519								263	309.1C	309.2C	315.1C							
345	Irish	H3*	129										263	309.1 C	315.1 C								
346	Irish	H3a	239G	519									151	152	263	309.1C	315.1C						
347	Irish	H3b3	129	519									263	309.1C	309.2C	315.1C							
348	Irish	H4*	519										152	263	309.1C	315.1C							
349	Irish	H4*	Anderson										152	263	315.1 C								
350	Irish	H4*	Anderson										263	315.1 C									
351	Irish	H45b	519										263	315.1 C									
352	Irish	H5	304										263	315.1C	456								
353	Irish	H51	278	519									73	263	309.1C	315.1C							
354	Irish	H53	519										93	263	309.1 C	315.1 C							
355	Irish	H5a	304										263	315.1 C									
356	Irish	H5a2	111	304									200	263	315.1C	456	523 del AC						
357	Irish	H6*	362	482									239	263	309.1C	315.1C							
358	Irish	H6a1*	129	153	362								152	239	263	309.1 C	315.1 C						
359	Irish	H7*	299										152	263	315.1 C								
360	Irish	H7*	266G	519									263	309.1 C	315.1 C								
361	Irish	H7*	168										263	315.1 C									
362	Irish	H7a1	261										263	315.1 C									
363	Irish	HV0	298										72	263	309.1C	315.1C							

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)										HVII									
364	Irish	HV0*	129	298									72	195	263	309.1 C	315.1 C					
365	Irish	I	129	223	391								73	199	204	250	263	315.1 C				
366	Irish	I*	129	223	391								73	146	152	199	204	207	250	263	315.1 C	
367	Irish	I*	129	223	391								73	152	199	204	207	250	263	315.1 C		
368	Irish	I1a1	129	172	223	311	362	391					73	199	203	204	250	263	315.1 C			
369	Irish	J1	69	126									73	263	295	315.1C	462	489				
370	Irish	J1b1a1a	69	126	145	172	192	222	261	519			73	242	263	295	315.1C	462	489			
371	Irish	J1b1a1a	69	126	145	172	192	222	261				73	242	263	295	315.1 C					
372	Irish	J1b1a1a	69	126	145	172	192	222	257	261			73	242	263	295	315.1 C					
373	Irish	J1b1a1a	69	126	145	172	192	222	261				73	242	263	295	315.1C	462	489			
374	Irish	J1c	69	126									73	185	263	295	315.1 C					
375	Irish	J1c	69	126									73	228	263	295	315.1 C					
376	Irish	J1c	63	69	93	126							73	228	263	295	315.1 C					
377	Irish	J1c	69	126	342								73	151	152	185	228	263	295	315.1 C		
378	Irish	J1c*	69	126	300								73	185	228	263	295	309.1 C	315.1 C			
379	Irish	J1c2c1	69	126	183C	189	519	527					73	185	188	222	228	263	295	315.1C	462	489
380	Irish	K*	224	311	519								73	146	152	263	309.1 C	315.1 C				
381	Irish	K1a*	224	311	519								73	146	195	263	309.1 C	315.1 C	497	524 Ins AC		
382	Irish	K1a*	93	224	311	519							8	73	195	263	315.1 C	497	524.1 Ins ACACAC			
383	Irish	K1a*	93	224	311	519							52	73	195	263	315.1C	497				
384	Irish	K1a10a	48	224	291	311							73	195	263	315.1 C						
385	Irish	K1a10a	48	224	291	311	519						73	152	195	263	315.1C	497	524.1 In:	523.2A		
386	Irish	K1a24a	145	224	311	519							73	150	152	195	263	315.1 C	497	523 del AC		
387	Irish	K1c	224	311	519								73	146	152	263	315.1 C	498 del				
388	Irish	K1c	224	311	519								73	146	152	263	315.1C	498 del	523 del AC			
389	Irish	K2b1a	224	270	311								73	146	263	309.1 C	315.1 C					
390	Irish	T	126	294									73	263	309.1 C	315.1 C						
391	Irish	T	126	294									73	263	315.1 C							
392	Irish	T*	126	189	294	519							73	152	263	315.1C						
393	Irish	T1a*	126	163	186	189	294						73	152	195	239	263	309.1 C	315.1 C			
394	Irish	T1a*	126	163	186	189	294						73	152	183	195	263	315.1 C				
395	Irish	T2b	126	294	296	304	519						73	263	309.1C	315.1C						
396	Irish	T2b	126	265 C	294	296	304						73	263	309.1 C	315.1 C						
397	Irish	T2b21	126	294	304	519							57	73	152	263	309.1C	315.1C	523 del AC			
398	Irish	T2b6	126	294	296	304	519						73	263	315.1C	458						
399	Irish	T2e*	126	153	294								73	150	263	315.1 C						
400	Irish	T2f	126	189	294	296	519						73	152	263	315.1C						
401	Irish	U2e1	51	129C	183C	189	193.1	362	519				73	152	217	263	309.1C	309.2C	315.1C	340	508	524 Ins AC
402	Irish	U5a*	192	256	270								73	263	315.1 C							
403	Irish	U5a1*	93	220A/C	256	270	399						73	263	315.1C							
404	Irish	U5a1*	192	256	270	311	399						73	199	263	309.1 C	315.1 C					
405	Irish	U5a1*	192	239	256	270	319	399					73	150	152	263	315.1 C					
406	Irish	U5a1c1	192	256	270	320	399						73	153	195	263	315.1 C					
407	Irish	U5a2a	114 A	192	256	270	294						73	152	263	309.1 C	315.1 C					
408	Irish	U5b2a2	189	192	270	398	544						73	150	263	315.1C						
409	Irish	U6a3b	93	172	183C	189	219	278					73	151	185	263	315.1C					
410	Irish	V9a1	192	219	502								72	204	207	263	309.1C	309.2C	315.1C			
411	Irish	W1*	223	292									73	119	189	195	204	263	315.1 C			
412	Irish	W1b	93	223									73	189	195	204	207	227	263	309.1 C	315.1 C	
413	Irish	W5a	129	223	292	362							73	189	194	195	204	207	263	309.1 C	315.1 C	
414	Irish	W5a	129	223	292	362	519						73	189	194	195	204	207	234G/A	263	309.1 C	315.1 C
415	Asian	A11	223	258	290	293C	319						73	152	235	263	309.1C	315.1C	523 del AC			
416	Asian	A4	223	290	319	362							73	152	235	263	309.1C	315.1C				
417	Asian	B6a	129	145	182C	183C	189						73	150	185	263	309.1C	309.2C	315.1C			
418	Asian	B6a	93	129	145	182C	183C	189					73	150	263	309.1C	309.2C	315.1C				
419	Asian	D4*	223	291	362	390 G/A							73	263	315.1C	489						
420	Asian	D5a2a1	92	164	167	182C	183C	189	193.1C	223	266	362	519	73	150	195	199	263	315.1C	489	523 del AC	
421	Asian	G2a1d2	278	362									73	260	263	315.1C	489					
422	Asian	H*	519										263	309.1C	315.1C	318A	455.1T					
423	Asian	H*	291	519									263	309.1C	315.1C	523 del AC						
424	Asian	H2a*	114										263	309.1C	315.1C							
425	Asian	H2a1	114										263	309.1C	315.1C							
426	Asian	H5	304										263	309.1C	315.1C	456	523 del AC					
427	Asian	HV12b	129	304	356								263	309.1C	315.1C							
428	Asian	HV2	217	325									72	73	152	182	195	263	309.1C	315.1C	523 del AC	573.1C Het
429	Asian	J1d	69	126	193	519							73	152	195	263	295	315.1C	462	489	524 ins AC	
430	Asian	J2b1a*	69	126	193	274	278						73	150	152	263	295	315.1C	489	523.1C	523.2A	
431	Asian	J2b1a2	69	126	193	278	519						73	150	152	195	235	263	295	315.1C	489	
432	Asian	K2a5	224	311	519								73	146	152	185	189	204	263	309.1C	315.1C	324
433	Asian	M*	75	92	157	223	399	519					73	152	263	315.1C	489	511				
434	Asian	M*	136	217	223	319	381						73	94	173	204	263	315.1C	482	489		
435	Asian	M*	223	239	311	519							73	150	263	309.1C	315.1C	316	489			
436	Asian	M*	189	193.1C	209	223	233	261	274	304	305	519	73	143	263	309.1C	315.1C	489	523 del AC			

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)											HVII												
437	Asian	M*	223	519										73	146	263	315.1C	489	524 ins ACAC							
438	Asian	M*	92	223	497	519								73	195	214	263	309.1C	315.1C	489							
439	Asian	M*	223	234	295G	301	311	519						73	263	315.1C	489										
440	Asian	M*	111	180	192	223	275	519						73	150	263	309.1C	309.2C	315.1C	489							
441	Asian	M*	153	223	286A	519								73	263	309.1C	309.2C	315.1C	489	573.1CCC							
442	Asian	M*	183C	189	209	223	274	278						73	234	263	309.1C	315.1C	489	523 del AC							
443	Asian	M*	92	223	362	519								73	263	309.1C	315.1C	489									
444	Asian	M*	145	223	300	316	519							73	146	152	263	309.1C	315.1C	489							
445	Asian	M12b	129	172	223	234	290	312	519					73	146	263	309.1C	315.1C	489								
446	Asian	M18a	172	223	311	318T	519							73	152	194	246	263	315.1C	489							
447	Asian	M2a1a	223	270	319	352	519							73	195	204	263	309.1C	315.1C	447G	489	523 del AC					
448	Asian	M2b1b	169.1C	189	223	235	274	319	320	519				73	152	182	195	263	309.1C	315.1C	447G	471	489	549	523 del AC		
449	Asian	M3*	126	223	311	519								73	203	204	217	263	309.1C	315.1C	482	489					
450	Asian	M3*	126	223	294	311	519							73	204	217	263	315.1C	482	489							
451	Asian	M3*	126	223	311	519								73	204	217	263	315.1C	482	489							
452	Asian	M3*	126	184	223	519								73	263	315.1C	317	482	489								
453	Asian	M3*	126	189	223	519								73	204	207	263	315.1C	482	489							
454	Asian	M3*	126	223	519									73	204	263	309.1C	315.1C	482	489	523 del AC						
455	Asian	M3*	126	184	223	390	519							73	204	207	263	315.1C	482	489							
456	Asian	M3*	126	180	223	445	519							73	185	263	315.1C	482	489								
457	Asian	M3*	126	223	233	266	344	519						73	204	207	263	315.1C	482	489							
458	Asian	M3*	126	223	311	519								73	204	217	263	309.1C	315.1C	482	489						
459	Asian	M3*	126	189	223	519								73	152	263	315.1C	482	489								
460	Asian	M3*	126	223	519									73	146	204	263	309.1C	315.1C	482	489						
461	Asian	M3*	126	223	519									73	203	204	263	309.1C	309.2C	315.1C	482	489					
462	Asian	M3*	126	147	223	319	519							73	263	315.1C	482	489									
463	Asian	M3*	126	223	519									73	263	315.1C	482	489									
464	Asian	M30*	179 del	223	302	519								73	152	195A	263	315.1C	489	523 del AC							
465	Asian	M30*	179 del	223	243	519								73	195A	263	309.1C	315.1C	489	523 del AC							
466	Asian	M30*	223	234	311	519								73	195A	263	309.1C	315.1C	489	523 del AC							
467	Asian	M30*	223	234	239	519								73	195A	263	315.1C	489	523 del AC								
468	Asian	M30*	223	234	519									73	195A	263	309.1C	315.1C	489	523 del AC							
469	Asian	M30*	223											73	195A	263	315.1C	489	523 del AC								
470	Asian	M30*	172	223	519									73	152	195A	200	263	309.1C	315.1C	489	523 del AC					
471	Asian	M33a2	169	172	223	519								73	263	309.1C	315.1C	462	489								
472	Asian	M35	223	519										73	199	263	315.1C	489									
473	Asian	M37e	111	184	189	223	295	519						73	146	195	198	263	309.1C	315.1C	489							
474	Indian	M39b1	129	223	304									55.1T	59delT	60delT	65.1T	66T	73	153	263	309.1C	315.1C	463	485	489	
475	Asian	M39b1	223	304										55.1T	59 del T	65.1T	66T	73	153	263	279 C/T	309.1C	315.1C	463	485	489	
476	Asian	M3a1	223	311	344	356	519							73	204	217	263	309.1C	309.2C	315.1C	482	489	524.1 Ins AC Het				
477	Indian	M3c1	51	183C	189	223	294	519						73	152	263	309.1C	309.2C	315.1C	482	489	523-4del ac					
478	Asian	M3c1	51	183C	189	223	294	519						73	152	263	309.1C	315.1C	482	489	523 del AC						
479	Asian	M3c1b	179	183C	189	193.1C	223	294	519					73	146	152	245	263	315.1C	482	489	523 del AC					
480	Asian	M4	145	176	223	261	311	519						73	263	315.1C	485	489	Heteroplasmy after 523 insert CA								
481	Asian	M4*	145	176	223	261	311	519						73	263	315.1C	485	489									
482	Asian	M4*	145	176	223	224	261	311	319	519				73	263	309.1C	315.1C	489									
483	Asian	M4*	93	145	176	223	261	311	519					73	152	263	315.1C	489									
484	Asian	M4*	38	86	145	223	311	519						73	263	315.1C	489										
485	Asian	M40a1a	179	223	289	294	319	356	463					73	125	127	152	200	204	249	263	315.1C	489				
486	Asian	M5*	129	223	519									73	263	309.1C	309.2C	315.1C	489								
487	Asian	M5a2a1a	129	223	265C	311	519							73	263	315.1C	489										
488	Asian	M5b2	48	129	218	223								57A	73	194	263	309.1C	309.2C	315.1C	489	523 del AC					
489	Asian	M5c1	129	182C	183C	189	193.1	223	468	519				73	263	309.1C	309.2C	315.1C	489	523 del AC							
490	Asian	M65a	223	234	289	519								73	152	263	309.1C	315.1C	489	511							
491	Asian	M65a	223	289										73	185	263	309.1C	315.1C	489	511							
492	Asian	M65b	223	311	319	519								73	241	263	309.1C	315.1C	489	511							
493	Asian	M65b	223	311	519									73	241	263	315.1C	372.1T	489	511							
494	Asian	M6a	177	188	223	231	362	519						73	146	263	309.1C	315.1C	461	489							
495	Asian	M6a	111	223	231	362	390	519						73	263	315.1C	461	489									
496	Asian	M6b	184	223	256G	362								73	146	152	263	309.1C	315.1C	461	489	523 del AC					
497	Asian	M71*	129	140	223	271								73	143	146	151	263	315.1C	489							
498	Asian	M7b*	129	189	192	223	297	519						73	150	199	263	309.1C	315.1C	489							
499	Asian	N1a1a	147A	172	223	248	320	355	519					73	152	199	204	263	309.1C	315.1C	573.1CC						
500	Asian	N21	182	193	223	519								73	150	195	263	315.1C	337 del A								
501	Asian	N8	95	263	274	311	343	357	519					73	152	199	263	315.1C									
502	Asian	R2	71	519										73	152	263	309.1C	315.1C									
503	Asian	R30*	172	183C	189	356	519							73	131	204	207	263	309.1C	315.1C							
504	Asian	R30a	126	148	209	362	519							73	152	263	309.1C	315.1C									
505	Asian	R30a	51	184										73	263	309.1C	315.1C	481A	523.1CACA								
506	Asian	R30b1	183C	189	194C	195	298	299						73	143	263	299 del C	309.1C	315.1C	373	523 del AC						
507	Asian	R31a	172	304	519									73	146	263	315.1C	315.2C	338	522 del CA							
508	Asian	R31a1	172	304	362	463	519							73	146	152	263	315.1C	315.2C	338	523 del AC						
509	Asian	R5a2b	136	158	174	248	266	304	325	356	399	519	524 A/G	73	152	263	315.1C	523 del AC									

Table 4.1 Mitochondrial haplogroups and control region sequences for 537 individuals

Sample No.	Ethnicity	Haplogroup	HVI (16,xxx)								HVII								
510	Asian	R6a*	129	179	227	245	266	278	362	519	73	94	195	246	263	309.1C	315.1C	523 del AC	
511	Asian	R6a*	179	227	245	266	278	362	519		73	152	195	246	263	309.1C	315.1C	522 del CA	
512	Asian	R6a*	179	227	245	266	278	362	519		73	152	195	246	263	309.1C	315.1C	372	522 del CA
513	Asian	R6a1	213	362	519						73	143	228	263	309.1C	315.1C			
514	Asian	R7*	187	241T	319	342	519				73	152	263	309.1C	315.1C				
515	Asian	R7*	260	261	319	362					73	263	309.1C	315.1C	480	524 ins AC			
516	Asian	R8a1b	93	172	519						73	195	263	309.1C	315.1C				
517	Asian	T1a*	126	163	186	189	294	519			73	152	195	263	309.1C	315.1C			
518	Asian	T2*	126	275	294	296	325	519			73	195	263	309.1C	309.2C	315.1C	452		
519	Asian	T2*	126	172	294	296	325	519			73	195	263	309.1C	315.1C				
520	Asian	T2g1	126	294	296						73	146	200	263	315.1C	523 del AC			
521	Asian	U1*	182C	183C	189	234	249	519	527		73	263	285	309.1C	315.1C				
522	Asian	U2a	51	93	206C	230	311	319	456	519	73	150	185	189	263	309.1C	309.2C	315.1C	573.1CCC
523	Asian	U2b1	51	168	355	519	527				73	146	152	185	189 A/G	263	315.1C		
524	Asian	U2b1	51	168	311	519					73	146	263	315.1C	523del AC				
525	Asian	U2b1	51	168	320						73	146	263	309.1C	315.1C				
526	Asian	U2b1	51	168	311	519					73	146	263	315.1C	522 del CA				
527	Asian	U2b1	51	168	172	234	287	359			73	146	185	263	309.1C	315.1C	522 del CA		
528	Asian	U2b1	51	168	311	519					73	146	153	263	309.1C	315.1C	522 del CA		
529	Asian	U2b1	51	168	185	224	242	399	519		73	146	152	195	263	309.1C	315.1C		
530	Asian	U2c	51	126	178	179	234	247	318 A/G		73	146	152	263	315.1C	573 del			
531	Asian	U4a1	51	134	356	519					73	152	195	236	263	309.1C	315.1C	499	
532	Asian	U7	189	309	318T	519					73	152	263	309.1C	315.1C	523 del AC			
533	Asian	U7a	309	318T	519						73	151	152	263	309.1C	315.1C	522 del CA		
534	Asian	U7a	309	318C	519						73	151	152	263	309.1C	315.1C	523 del AC		
535	Asian	U7a4	126	207	292	309	318T	519			73	151	152	263	309.1C	315.1C	523 del AC		
536	Asian	Z3a	150	185	223	260	298	519			73	152	207	249 del A	263	309.1C	309.2C	315.1C	489
537	Asian	Z3a	150 C/T	185	223	260	298	519			73	152	207	249 del A	263	309.1C	315.1C	489	

Sequences are expressed as changes from the rCRF. Those in the mitochondrial region 16024-16569, which includes HVI, are listed without the 16,000 prefix. All changes are transitions (i.e. C/T or A/G) unless explicitly stated. Sample numbers are colour coded to represent sequence coverage, clear indicates that only HVI and HVII were sequenced, light pink indicates that the entire control region was sequences, the bright cerise pink denotes that the entire mitochondrial genome was sequenced while red indicates that only part of the coding region was sequenced in addition to the control region. Sequence changes highlighted in yellow indicate diagnostic changes from R* with respect to haplogroup classification.

Table 4.2 Mitochondrial coding region sequencing results

Sample No.	Ethnicity	Haplogroup	3010 G-A	3594 C-T	4216 T-C	7028 C-T	10398 A-G	10400 C-T	12372 G-A	12705 C-T	4769 A-G	3915 G-A	3992 C-T	4336 T-C	4745 A	4793 A-G	6776 T-C	Coding						
20	Jamaican	L3b*		C			G	C																
21	Jamaican	L3b*		C			G	C																
25	Jamaican	L3d*																7424						
28	Jamaican	L3d4		C			G	C																
45	Barbadian	H1c13					A	C	G	C								750	1438	3010	4769	8645	8860	9965	15326
46	Barbadian	HV0*				T			G	C														
48	Barbadian	L1b				T			G															
77	Barbadian	L3b*		C			G	C																
102	Afro-Caribbe	K1a	G	C	T	T	G	C	A	C														
103	Afro-Caribbe	L0a1*	G	T	T	T	G	C	G	T														
106	Afro-Caribbe	L1b	G	T	T	T	G	C	G	T														
107	Afro-Caribbe	L1b	G	T	T	T	G	C	G	T														
108	Afro-Caribbe	L1b*	G	T	T	T	G	C	G	T														
123	Afro-Caribbe	L1c*	G	T	T	T	G	C	G	T														
125	Afro-Caribbe	L1c1a1	G	T	T	T	A	C	G	T														
126	Afro-Caribbe	L1c2b	G	T	T	T	G	C	G	T														
130	Afro-Caribbe	L1c3a1b	G	T	T	T	G	C	G	T														
136	Afro-Caribbe	L2a1*	A	T	T	T	G	C	G	T														
137	Afro-Caribbe	L2a1*	G	T	T	T	G	C	G	T														
152	Afro-Caribbe	L2a1a2	G	T	T	T	G	C	G	T														
162	Afro-Caribbe	L2c*	G	T	T	T	G	C	G	T														
171	Afro-Caribbe	L3b*	G	C	T	T	G	C	G	T														
172	Afro-Caribbe	L3b*		C			G	C																
175	Afro-Caribbe	L3b*		C			G	C																
183	Afro-Caribbe	L3d*		C			G	C																
185	Afro-Caribbe	L3d1a		C			G	C																
199	Afro-Caribbe	L3e2*		C			G	C																
209	Afro-Caribbe	L3e2b*	G	C	T	T	G	C	G	T														
213	Afro-Caribbe	L3e3		C			G	C																
220	Afro-Caribbe	L3f1a		C			G	C										4218	4350	5601				
221	Afro-Caribbe	L3f1b*	G	C	T	T	G	C	G	T														
231	Afro-Caribbe	L3k		C			G	C																
232	Afro-Caribbe	M1a1	G	C	T	T	G	T	G	T														
233	Afro-Caribbe	M65a	G	C	T	T	G	T	G	T														
234	Afro-Caribbe	M6a	G	C	T	T	G	T	G	T														
235	Afro-Caribbe	U8a1	G	C	T	T	A	C	A	C														
238	Caucasian	H*																750	1438	3505	4769	8860	13748	15326
239	Caucasian	H*	G	C	T	C	A	C	G	C		G	C	T	A	A	T	750	1438	4491	4769	5147	8860	13194	15326
240	Caucasian	H*	G	C	T	C	A	C	G	C	G	G	C	T	A	A	T	750	1438	1653	1719	4769	8269	8860	15326
241	Caucasian	H1	A						G			G	C	T	A	A	T								
242	Caucasian	H1*	A			C					G	G	C	T	A	T	T								
243	Caucasian	H1*	A			C					G	G	C	T	A	T	T								
244	Caucasian	H1*																750	1438	3010	4769	8251	8860	9960	11506	15326
245	Caucasian	H1*	A	C	T	C	A	C	G	C																
246	Caucasian	H1*	A	C	T	C	A	C	G	C		G	C	T	A	A	T								
247	Caucasian	H1*	A	C		C	A	C	G	C																
248	Caucasian	H1*	A	C	T	C	A	C	G	C																
249	Caucasian	H1*	A			C						G	C	T	A	A	T								
250	Caucasian	H1*	A			C						G	C	T	A	A	T								
251	Caucasian	H1*	A			C						G	C	T	A	A	T								
252	Caucasian	H1*	A			C						G	C	T	A	A	T								
253	Caucasian	H1*	A			C						G	C	T	A	A	T								
254	Caucasian	H1*	A			C						G	C	T	A	A	T								
255	Caucasian	H1*	A			C			G		G	G	C	T	A	A	T								
256	Caucasian	H10e				C												14470							
257	Caucasian	H11a																13759							
258	Caucasian	H11a1	G			C						G	C	T	A	A	T								
259	Caucasian	H15a1																							
260	Caucasian	H16c	G			C					G	G	C	T	A	T	T	750	*1438*	*4769*	*8860*	*9071*	*10394*	*15326*	*15475*
261	Caucasian	H17c																750	1438	3915	4769	8860	12397	15326	
262	Caucasian	H1a1	A	C	T	C	A	C	G	C		G	C	T	A	A	T								

Table 4.2 Mitochondrial coding region sequencing results

Sample No.	Ethnicity	Haplogroup	3010	3594	4216	7028	10398	10400	12372	12705	4769	3915	3992	4336	4745	4793	6776	Coding										
263	Caucasian	H1b	A	C	T	C	A	C	G	C																		
264	Caucasian	H1ba	A			C					G	G	C	T	A	A	T											
265	Caucasian	H1c1	A	C	T	C	A	C	G	C																		
266	Caucasian	H24a	G			C						G	C	T	A	A	T	750	1438	3333	4769	8860	13327	15326				
268	Caucasian	H2a*	G	C	T	C	A	C	G	C		G	C	T	A	A	T	750	5206	8860	15326							
269	Caucasian	H2a1	G			C					A	G	C	T	A	A	T											
270	Caucasian	H2a1d	G			C					G	G	C	T	T	A	T	750	951	4216	4769	6260	8679	8860	15326			
271	Caucasian	H2a1f				C			G		A					A												
272	Caucasian	H3*										G	C		A		C											
273	Caucasian	H3*										G	C		A		C											
274	Caucasian	H3*										G	C		A		C											
275	Caucasian	H3*	G			C			G			G	C	T	A	A	C											
276	Caucasian	H3*	G	C	T	C	A	C	G	C		G	C	T	A	A	C											
277	Caucasian	H3as	G			C						G	C	T	A	A	C	750	1438	4769	6776	8860	9316	15326				
278	Caucasian	H4	G	C	T	C	A	C	G	C		G	T		A		T											
281	Caucasian	H51	G	C	T	C	A	C	G	C	G	G	C	T	A	A	T	750	1438	4769	8860	11440	15326					
282	Caucasian	H6a*	G			C			G			A	C	T	A	A	T											
283	Caucasian	H6a*	G	C	T	C	A	C	G	C		A	C	T	A	A	T											
284	Caucasian	H6a*	G			C						A	C	T	A	A	T											
285	Caucasian	H6a1b2	G	C	T	C	A	C	G	C								750	1438	2887	3915	4727	4769	8860	9254	9380	10589	15326
286	Caucasian	H7*	G			C					G	G	C	T	A	G	T											
287	Caucasian	HV0*	G	C	T	T	A	C	G	C																		
288	Caucasian	HV0*	G	C	T	T	A	C	C	G																		
289	Caucasian	HV16	G			T						G	C	T	A	A	T	750	1438	2339	2706	4123	4769	5581	7028	8860	12492T	14180	15326
290	Caucasian	HV5	G			T						G	C	T	A	A	T	750	1438	2706	4769	7028	8860	12133	13105	15326			
291	Caucasian	I*	G	C	T	T	G	C	G	T																		
293	Caucasian	I1a1	G	C	T	T	G	C	G	T																		
294	Caucasian	J1b1a	A	C	C	T	G	C	G	C																		
295	Caucasian	J1b1a1	A	C	C	T	G	C	G	C																		
296	Caucasian	J1b1a1a	A	C	C	T	G	C	G	C																		
297	Caucasian	J1c1	A	C	C	T	G	C	G	C																		
298	Caucasian	J1c2	A	C	C	T	G	C	G	C																		
300	Caucasian	K1a	G	C	T	T	G	C	A	C																		
301	Caucasian	K1a*	G	C	T	T	G	C	A	C																		
302	Caucasian	K1a*	G	C	T	T	G	C	A	C																		
303	Caucasian	K1a1	G	C	T	T	G	C	A	C																		
304	Caucasian	K1a4a1a2b	G	C	T	T	G	C	A	C																		
307	Caucasian	K1c	G	C	T	T	G	C	A	C																		
308	Caucasian	K1c					G	C	A	C																		
310	Caucasian	M2c	G	C	C	T	G	T	G	T																		
312	Caucasian	T1a	G	C	C	T	A	C	G	C																		
314	Caucasian	T2*	G	C	C	T	A	C	G	C																		
315	Caucasian	T2b	G	C	C	T	A	C	G	C																		
316	Caucasian	T2b	G	C	C	T	A	C	G	C																		
318	Caucasian	U5a1*				T			A																			
321	Caucasian	U5a2a				T			A																			
325	Irish	H*	G			C						G	C	T	A	A	T	750	1391	1438	4769	8860	10632 T/C	12358	15110	15326		
326	Irish	H*	G			C					G	G	C	T	A	A	T	750	1438	4769	8860	15326						
327	Irish	H1*	A			C						G	C	T	A	A	T											
328	Irish	H1*	A			C						G	C	T	A	A	T											
329	Irish	H1*	A			C						G	C	T	A	A	T											
330	Irish	H1*	A			C						G	C	T	A	A	T											
331	Irish	H1*	A			C						G	C	T	A	A	T											
332	Irish	H1*	A			C						G	C	T	A	A	T											
333	Irish	H1*	A			C						G	C	T	A	A	T											
334	Irish	H1*	A			C						G	C	T	A	A	T											
335	Irish	H1*	A			C						G	C	T	A	A	T											
336	Irish	H1a1	A			C						G	C	T	A	A	T											
337	Irish	H1b	A			C						G	C	T	A	A	T											
338	Irish	H1b*	A	C	T	C	A	C	G	C		G	C	T	A	A	T											
339	Irish	H1i2*	A			C						G	C	T	A	A	T	750	1438	1766G	2581	3010	4769	6237A	8860	12346	15326	*15718*
340	Irish	H2a2a	G			C					A	G	C	T	A	A	T											
341	Irish	H2a2a1	G			C					A	G	C	T	A	A	T											

Table 4.2 Mitochondrial coding region sequencing results

Sample No.	Ethnicity	Haplogroup	3010	3594	4216	7028	10398	10400	12372	12705	4769	3915	3992	4336	4745	4793	6776	Coding
342	Irish	H3*	G			C						G	C	T	A	A	C	750 1438 4769 5132 6776 8860 15326
343	Irish	H3*	G	C	T	C	A	C	G	C		G	C	T	A	A	C	
344	Irish	H3*	G			C						G	C	T	A	A	C	
345	Irish	H3*	G			C						G	C	T	A	A	C	
347	Irish	H3b3	G			C	A	C				G	C	T	A	A	C	750 1438 2581 4769 4924 6776 8860 15326
348	Irish	H4*	G			C						G	T	T	A	A	T	
349	Irish	H4*	G			C						G	T	T	A	A	T	
350	Irish	H4*	G			C						G	T	T	A	A	T	
351	Irish	H45b	G			C						G	C	T	A	A	T	750 1438 4164 4769 8843 8860 12130 15326
353	Irish	H51	G			C						G	C	T	A	A	T	750 1438 4769 8860 11440 11887 15326
354	Irish	H53	G			C						G	C	T	A	A	T	750 1438 3450 4769 8860 9380 11167 15326
355	Irish	H5a	G			C						G	C	C	A	A	T	
358	Irish	H6a1*	G			C						A	C	T	A	A	T	
359	Irish	H7*	G			C						G	C	T	A	G	T	
360	Irish	H7*	G			C						G	C	T	A	G	T	
361	Irish	H7*	G			C						G	C	T	A	G	T	
362	Irish	H7a1	G			C						G	C	T	A	G	T	
369	Irish	J1	A	C	C	T	G	C	G	C								
370	Irish	J1b1a1a	A	C	C	T	G	C	G	C								
371	Irish	J1b1a1a	A	C	C	T	G	C	G	C								
372	Irish	J1b1a1a	A	C	C	T	G	C	G	C								
374	Irish	J1c	A	C	C	T	G	C	G	C								
375	Irish	J1c	A	C	C	T	G	C	G	C								
376	Irish	J1c	A	C	C	T	G	C	G	C								
377	Irish	J1c	A	C	C	T	G	C	G	C								
378	Irish	J1c*	A	C	C	T	G	C	G	C								
383	Irish	K1a*	G	C	T	T	G	C	A	C								
386	Irish	K1a24a	G	C	T	T	G	C	A	C								
388	Irish	K1c	G	C	T	T	G	C	A	C								
389	Irish	K2b1a	G	C	T	T	A	C	A	C								
390	Irish	T	G	C	C	T	A	C	G	C								
391	Irish	T	G	C	C	T	A	C	G	C								
392	Irish	T*	G	C	C	T	A	C	G	C								
393	Irish	T1a*	G	C	C	T	A	C	G	C								
394	Irish	T1a*	G	C	C	T	A	C	G	C								
395	Irish	T2b	G	C	C	T	A	C	G	C								
396	Irish	T2b	G	C	C	T	A	C	G	C								
397	Irish	T2b21	G	C	C	T	A	C	G	C								
398	Irish	T2b6	A	C	C	T	A	C	G	C								
400	Irish	T2f	G	C	C	T	A	C	G	C								
408	Irish	U5b2a2	G	C	T	T	A	C	A	C								
409	Irish	U6a3b	G	C	T	T	A	C	A	C								
417	Asian	B6a		C			A	C										750 1438 2706 3763 4769 5773 5893 5894C 5899.1 Ins CCCCCCCCC 7028 -9 del CCCCC 8860
417*																		9452 11719 11914 12950 13824 13928C 14305 14766 15326
418	Asian	B6a	G	C	T	T	A	C	G	C								11914 13824 13928C
419	Asian	D4*					G	T										3010 8860 14668 14766
422	Asian	H*	G			C					G	G	C	T	A	A	T	
423	Asian	H*	G			C					G	G	C	T	A	A	T	
424	Asian	H2a*	G		T	C					A	G	C	T	A	A	T	
425	South Asian	H2a1	G		T	C	A	C	G	T								750 951 8860 11252 15140 15326
427	Asian	HV12b	G	C	T	T	A	C	G	C								750 1284 1438 2706 4769 7028 8860 13889 15326 15355 15682
428	Asian	HV2	G	C	T	T	A	C	G	C								2706 7028 7193
430	Asian Afghan	J2b1a*																11204
431	Asian	J2b1a2																750 1438 2706 4216 4769 5633 6216 6893 7028 7476 8538 8860 8962 10172 10398
431*																		11251 11719 12612 12810 13708 13809A 14766 15212 15257 15326 15452A 15812
433	Asian	M*		C			G	T										
434	Asian	M*		C			G	T										
435	Asian	M*					G	T										
436	Asian	M*					G	T										
437	Asian	M*					G	T										
438	Asian	M*					G	T										
439	Asian	M*					G	T										
440	Asian	M*					G	T										

Table 4.2 Mitochondrial coding region sequencing results

Sample No.	Ethnicity	Haplogroup	3010	3594	4216	7028	10398	10400	12372	12705	4769	3915	3992	4336	4745	4793	6776	Coding
441	Asian	M*					G	T										
442	Asian	M*					G	T										10327 10398 10400
443	Asian	M*					A	C										*15043* 15326
444	Asian	M*					G	T										*6827 7028 8860 9095 9180*
464	Asian	M30*	G	C	T	T	G	T	G	T								
465	Asian	M30*		C			G	T										
466	Asian	M30*	G	C	T	T	G	T	G	T								
472	Asian	M35	G	C	T	T	G	T	G	T								
474	Indian	M39b1	G	C	T	T	G	T	G	T								
476	Asian	M3a1																4580 4703 4769 8860
477	Indian	M3c1	G	C	T	T	G	T	G	T								
484	Asian	M4*		C			G	T										*12007*
485	Asian	M40a1a	G	C	T	T	G	T	G	T								
486	British Pakist	M5*	G	C	T	T	G	T	G	T								*1822* 1888 2706 7028
489	Asian	M5c1																*750 1438 1888 2706 4769 4851 5319 5933T 6218 6413 6917 7028 7055 8065 8701*
489*																		*8790 8860 9540 10398 10400 10873 11719 12366 12705 14693 14766T 14783 15043 15301* 15326
496	Sri Lanka	M6b	G	C	C	T	G	T	G	T								
497	Asian	M71*																*15301 15326 15458*
500	Asian	N21	G	C	T	T	A	C	G	T								
503	Asian	R30*	G	C	T	T	A	C	G	C								750 *1438 1719* 2385G 2706 2906 *4596 4769* 7028 *8584 8860*
503*																		*10005 10328 11611 11719 12362 13302 14766* 15326 *15613*
504	Asian	R30a	G	*	T	T	A	C	G	C								750 1438 2056 2706 3158.1T 3316 4225 4232 4769 5237 5442 6764 7028 7274
504*																		8584 8860 9156 9242 9966 10646 11047A 11506 11719 12714 14766 15055 15326
505	Asian	R30a	G	C	T	T	A	C	G	C		G	C					2056 3316 3320 *8281-8289del* 8584 11047A *11152*
510	Asian	R6a*		C			A	C										*750 1438 2706 4769 4991 7028 7364 7984 8860 9254 11719 11776 12285* 13812 *14173 14180 14766* 15326
511	Asian	R6a*																*750 1438 2706 3535 4769 4991 7028 7364 7984* 8860 *9254 11719 11776 12285* 13812 *14766* 15326
512	Asian	R6a*		C			A	C										*2706 11719 11776*
513	Asian	R6a1																750 1438 2706 4769 5894 7028 7897 8860 11075 11719 12133 12285 14058 14225 14766 15067 15202 15326
514	Asian	R7*	G	C	T	T	A	C	G	C								*750 994 1119 1438 1442 1676 2706* 4769 *6413 7028 8167 8572* 8860
514*																		*9051 9110 10256 11464A 11719 12435A 13105 14131 14233 14766* 15326 *15924*
515	Asian	R7*																
516	Asian	R8a1b																709 750 1438 2706 2755 3384 4769 5510 5911 7028 7759
516*																		8860 9377 9449 9732A 11719 13212 13215 13782 14041 14766
518	Asian	T2*	G	C	C	T	A	C	G	C								
521	Asian	U1*				T	A	C	A	C								
524	Asian	U2b1	G	C	T	T	A	C										*3915 4093 11467*
525	Asian	U2b1																*3732 3849 3915 4093 11467*
526	Asian	U2b1																*3915 4093 11467*
527	Asian	U2b1																*3915 4093 11467 13651*
528	Asian	U2b1																*3915 4093 11467*
529	Asian	U2b1	G	C	T	T	G	C	A	C								750 *1438 1811* 3915 4093 4769 5186 7028 7115 8860 11204
529*																		11467 11719 12106 12308 12372 13194 13535 13708 14766 15049 15326
530	Asian	U2c																14935 *15061 15326*
534	Asian	U7a	G	C	T	T	A	C	A	C								12308 12372 12373
535	Asian Pakista	U7a4	G	C	T	T	A	C	A	C								

Sequence changes are listed here for samples that underwent additional coding region SNP analysis or sequencing. Results from the pyrosequencing SNPs are collated on the left while sequence changes observed with full or partial coding region analysis are displayed on the right. Colour coding is identical to that in Table 4.1, but with the addition of sequence changes highlighted in purple that indicate bases that are not present on the known mitochondrial phylogeny for that branch – some of these changes are observed in more than 1 individual (e.g. some of the R6* changes) indicating that the known mitochondrial phylogenetic tree of worldwide variation can be updated with this data. Sample numbers suffixed by an asterix indicate that the coding region sequence takes continues onto a second row.

Different branches of this phylogenetic tree can be given different names - these names are termed haplogroups. In mitochondrial DNA, the major branches are given haplogroup names with different letters of the alphabet. For example, one major branch found at a high frequency within Western Europe is named haplogroup H. Each sub-branch within this haplogroup is further named, for example an early mutation within haplogroup H occurred at base 3010 where a G mutated into an A, all individuals descended from the haplogroup H ancestor with this 3010 mutation belong to haplogroup H1. Further sub-branching continues the alternating nomenclature between letters and numbers, e.g. refinement of this H1 haplogroup could lead to H1b and then further to H1b3 etc.

4.1 Caribbean Populations

Displayed in Figure 4.1 is the most parsimonious phylogenetic tree that can be generated from the control region sequencing data of the sampled Jamaican population. Each node represents a specific mitochondrial control region sequence, and the size of the node is proportional to the number of individuals in the sampled population with that same sequence (in this data set all but one sequence is unique). The distance between nodes is proportional to the number of mutational steps needed to change from one sequence to another. Due to the availability in the public domain of mitochondrial sequencing data produced in the last 15 years, the major recent (i.e. *Homo sapiens*) evolutionary events in the mitochondrial genome have been reconstructed. This has enabled a fairly detailed phylogenetic tree to be created sequentially listing these mutational changes, and can (in a best case scenario) allow sequence data from new samples to be unambiguously attributed to a lineage (haplogroup). This haplogroup designation can be achieved due to the presence of specific changes within the sequence, e.g. haplogroup T can be assigned by characteristic changes in the HVI region at bases 16126 and 16294. The assignment of a haplogroup to a lineage doesn't mean that the sequence will match exactly to that in the known mitochondrial phylogenetic tree, just that it is known to have arisen from that evolutionary mitochondrial branch. Sometimes the sequence generated from HVI/HVII or the control region is sufficient to provide very fine haplogroup

assignment (e.g. L1c1a1) while at other times it is not possible to deduce any haplogroup affiliation without typing some or all of the coding region.

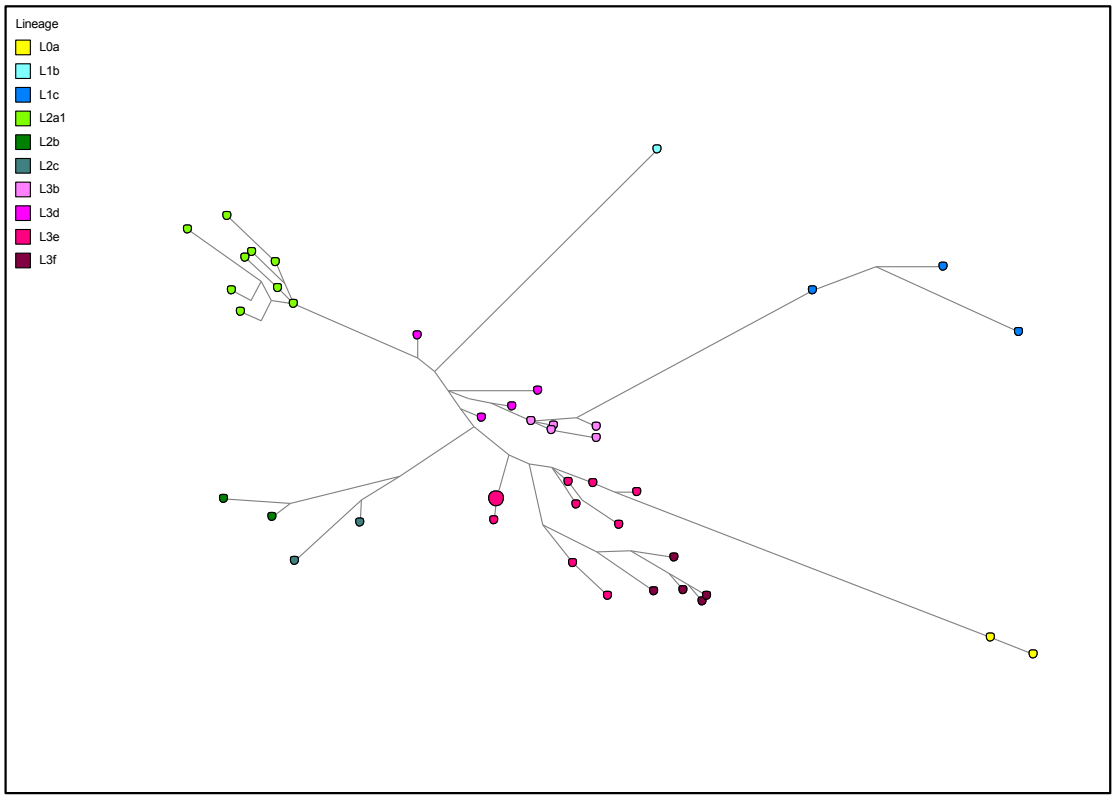


Figure 4.1 Phylogenetic tree of 44 Jamaican mitochondrial control region sequences
The size of circle is relative to the number of samples represented with that sequence (all bar one are single samples here) and the length of line between the nodes is proportional to the number of sequence differences between the samples. Node colour is a reflection of the L haplogroup subtype.

As Table 4.1 shows, all the Jamaican samples can be assigned to the African mitochondrial Haplogroup L, with most falling into the L3 subgroup. Figure 4.1 displays the different L subgroups in different colours, and while most samples within an L subtype cluster together, there is one L1 and four L2 sequences that do not cluster with the rest of their L subgroup based on the control region sequences alone; these relate to an L1b sequence clustering away from the three L1c sequences and L2b/c sequences clustering away from the remaining nine L2a1 sequences. Given the control region changes in these samples and the known evolutionary mutational events leading to the formation of these L subgroups, the classification for these samples is sound. The addition of a set of control region SNPs to the data (inferred from the known haplogroup if not actually typed in the sample) results in a better

separation of the four L subgroups as seen in Figure 4.2 and further refinement within these subgroups is still easily apparent.

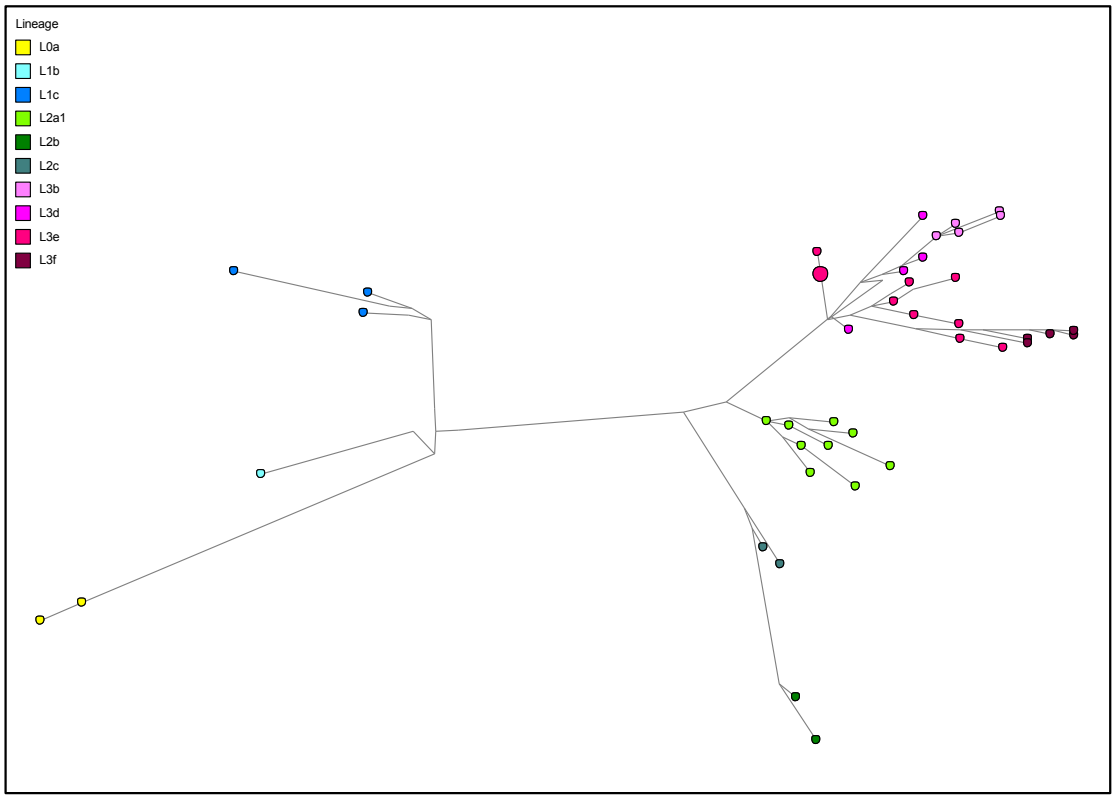


Figure 4.2 Phylogenetic tree of 44 Jamaican samples generated from control region sequences and data from 4 SNPs
As Figure 4.1 but with the addition of data on 4 SNPs (defining L0, L1, L2 and L3) that was either directly typed by pyrosequencing or full genome sequencing if haplogroup assignment was inconclusive, or inferred if a haplogroup had been securely designated on the basis of the control region sequence.

Figure 4.3 displays the results when the additional samples from Barbados and other assorted Caribbean Islands are added onto this skeleton. The vast majority of samples still fall within the L haplogroup, but the division between the subgroups shows some variation between the different islands with L1 and L2 predominating over L3 in the Barbadian population. In addition there are three Barbadian samples placing in haplogroups H, HV and U that are more often associated with European derived populations and hence take up outlier position on the tree (the three samples in the top right hand corner). One Trinidadian sample falls within haplogroup M which is seen widely across South Asia, and this also takes up an outlier position at the top centre of the tree. While these four samples falling into 'non-African' haplogroups do separate out from the rest of the sequences, the variation present in haplogroup L as

highlighted by the control region is so great that some individuals in the L1 and L0 subgroups are shown to be genetically more distant from L3 than these non-African sequences. This is not unexpected since the large variation present within the African haplogroup L is related to its status as the section containing the root of the phylogenetic tree (i.e. mitochondrial Eve), and has therefore been afforded more time for variations to accrue by virtue of being the oldest haplogroup.

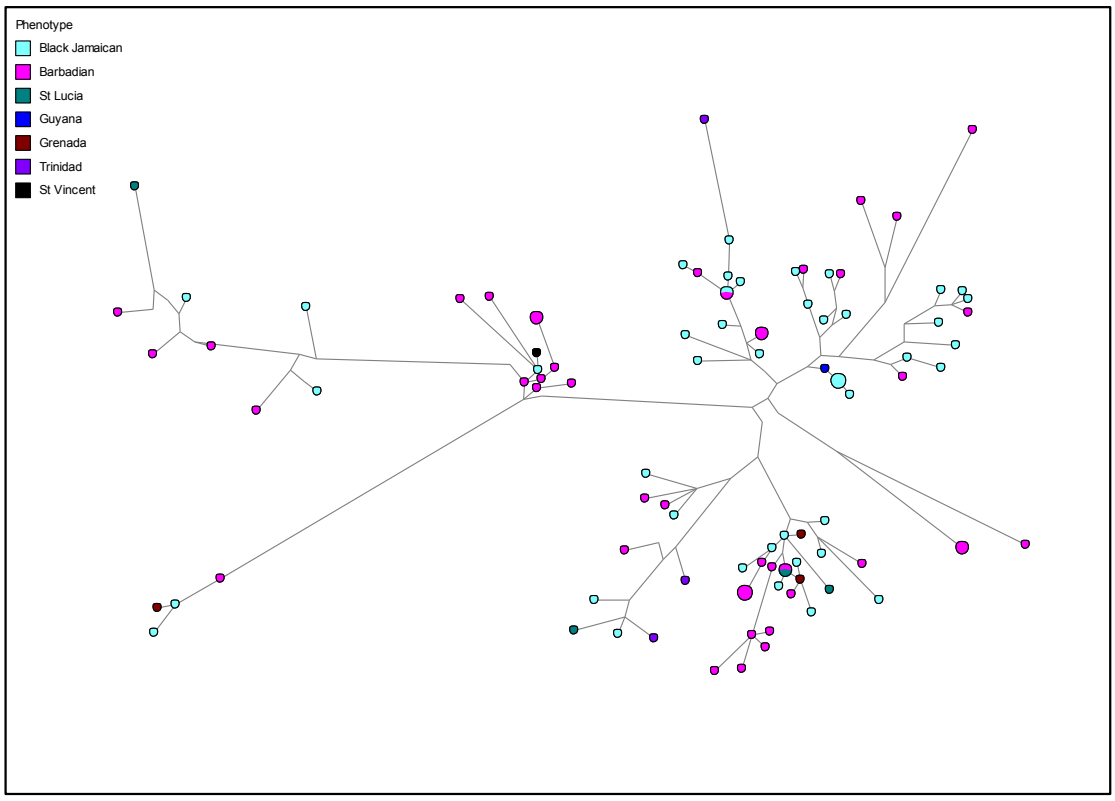


Figure 4.3 Phylogenetic tree of 44 Jamaican, 44 Barbadian and 12 other Caribbean mitochondrial control region sequences (and SNPs)
The addition of Barbadian and assorted Caribbean sequences onto the Jamaican skeleton depicted in Figure 4.2. Again, trees are generated from sequence data consisting of the entire mitochondrial control region plus inferred or typed information from key SNPs (the 4 SNPs defining the L sub-groups plus another 4 specific for haplogroups H, HV, U and M). The different populations are depicted in different colours.

Figure 4.4 provides a more in-depth look at the L3 haplogroup cluster where it is apparent that the Barbadian and Jamaican samples are both spread across the same L3b, L3d, L3e and L3f subtypes, however the frequency of the L3 haplogroup within the Jamaican population is noticeably higher than in the Barbadian (as seen in Figure 4.4 by the excess of blue dots). By contrast, Figure 4.5 and Figure 4.6 demonstrate that in haplogroups L1 and L2 there is more of a separation within the subtypes relative to the population.

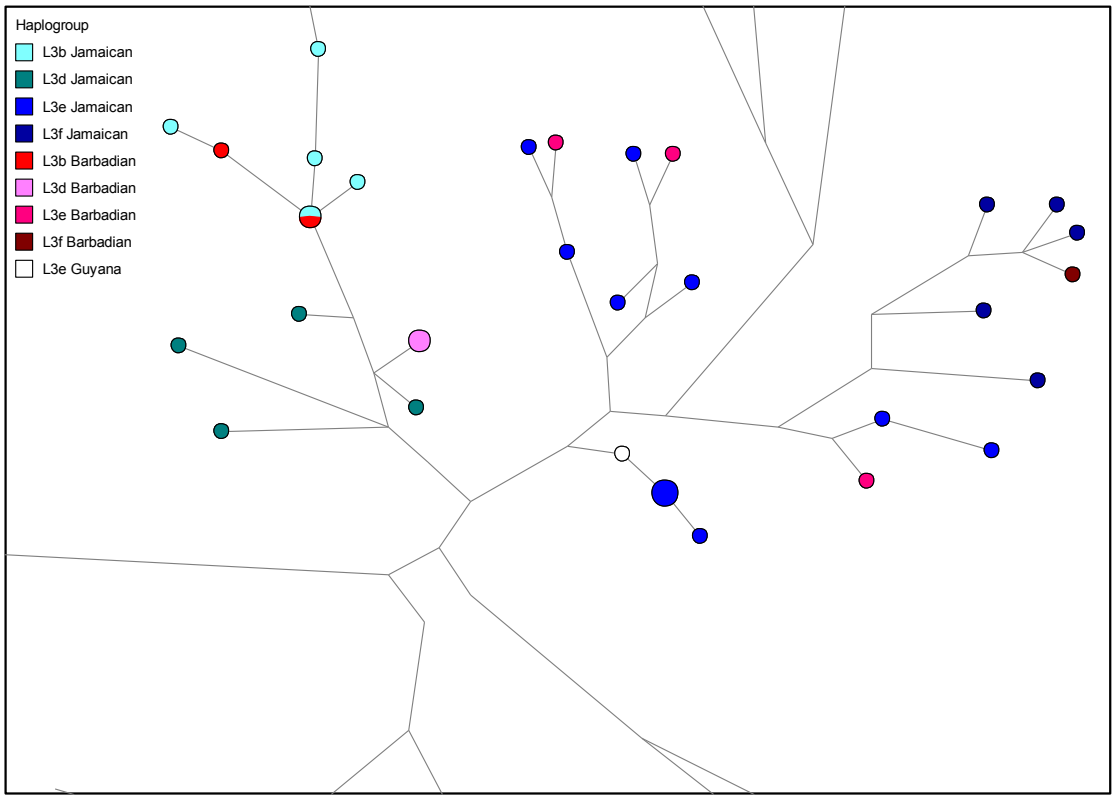


Figure 4.4 Magnified view of the L3 segment of the phylogenetic tree shown in Figure 4.3.
L3 subtype (as determined from the control region sequence), along with population, has been highlighted in this more detailed view.

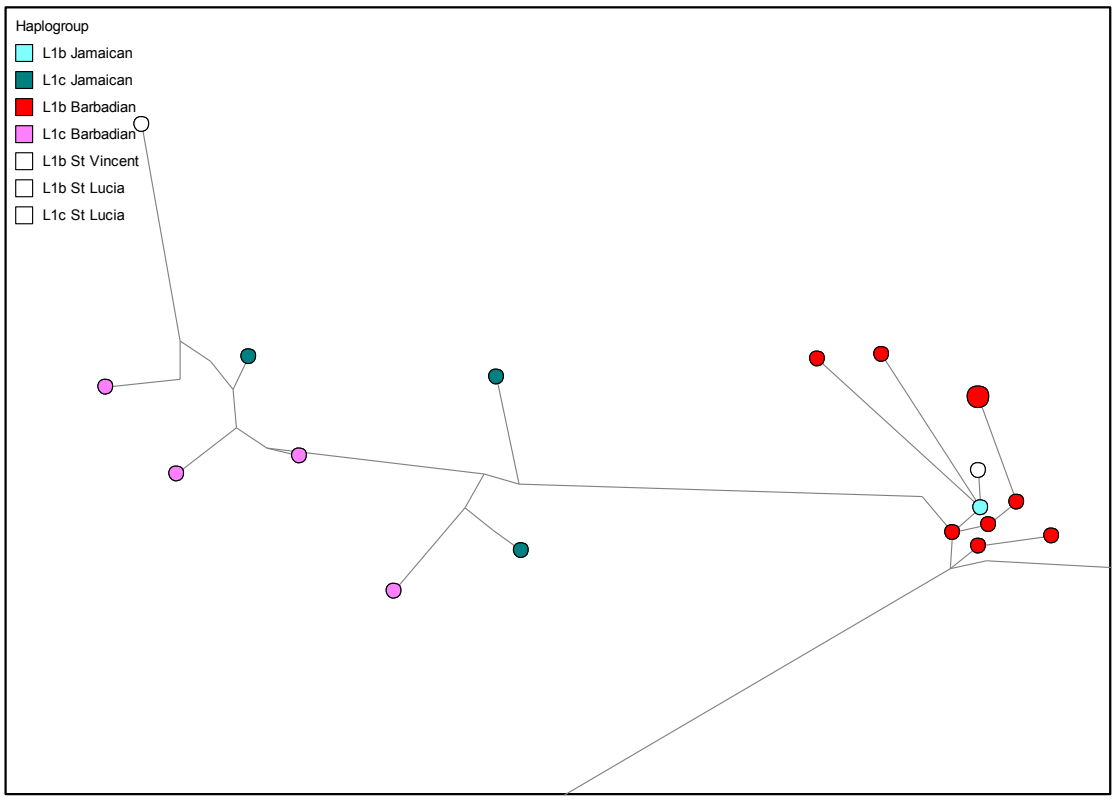


Figure 4.5 Magnified view of the L1 segment of the phylogenetic tree shown in Figure 4.3.
L1 subtype (as determined from the control region sequence), along with population, has been highlighted in this more detailed view.

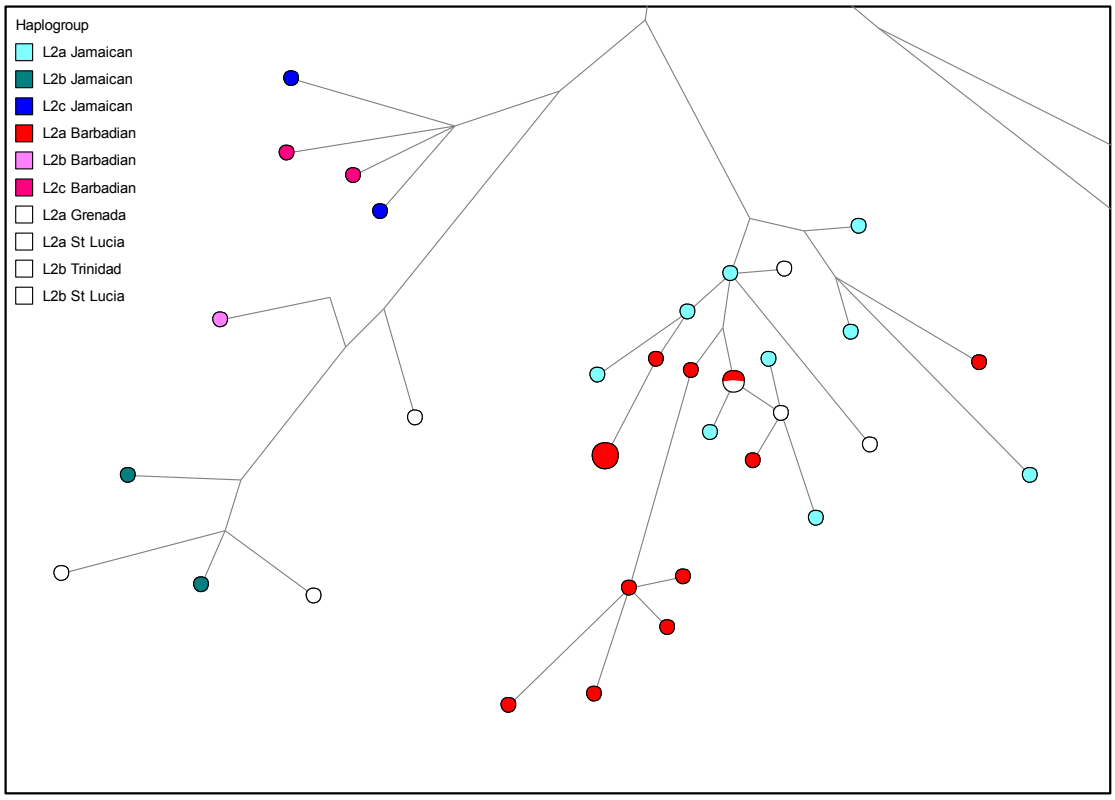


Figure 4.6 Magnified view of the L2 segment of the phylogenetic tree shown in Figure 4.3.
L2 subtype (as determined from the control region sequence), along with population, has been highlighted in this more detailed view.

Differing histories might explain why there is not a more uniform distribution of haplogroups across these two Caribbean islands. Barbados was settled by British colonists in 1627 and soon became a global force in the sugar industry requiring the influx of a large labour force [209]. White British workers first provided this labour either in the form of convicted prisoners or the poor working off their passage to the colony (indentured labour). During Cromwell's reign nearly 7000 Irish were also sent across. However it soon transpired that cheaper labour could be imported from West Africa through the slave trade and it's estimated that 387,000 slaves were shipped over to Barbados between 1627 and 1807, principally from tribes in what is now Ghana, but also from Nigeria. This resulted in changing demographics on the island where white Europeans made up 97% of the population in 1629 but only 21% by 1786. Unlike other Caribbean colonies where the European presence rapidly dwindled, this black/white (African/European) ratio remained about 3:1 throughout the 18th Century, with a significant poor white population present on the island. Additionally, from the start of the 18th century most black residents were born on the

island rather than arriving fresh from Africa, helping to shape a more defined Barbadian population. This is once again in contrast to other English speaking Caribbean islands where the mortality rate amongst the slaves was so great that a constant stream of new arrivals was needed [209]. Hence the presence of some more specific sequence clusters within the Barbadian samples that group separately from the Jamaican and UK Afro-Caribbean (see Figure 4.10) samples is not unexpected in relation to the historical context.

Previous analysis of a Jamaican population with autosomal markers shows a low level of European ancestry when compared with African American populations at just 6.8% [210]. This low level of admixture correlates with the mitochondrial results obtained here where all Jamaican samples contain mitochondria DNA that can be traced back to the African derived L haplogroup. The presence of three non-L lineages in the Barbados population that are more commonly found in White Europeans again is not unexpected given the history of the island. Even though these individuals classify themselves as Black Barbadian, it is apparent that at some point in the past they have a European female ancestor as the result of an admixture event between the African and European population groups present on the island.

4.2 UK Afro-Caribbean Population

Addition of the UK Afro-Caribbean samples onto this skeleton results in quite a complex picture (Figure 4.7). The overarching distribution is similar with most samples placing in haplogroups L1-L3, however it's apparent that there are many additional clusters appearing demonstrating that the Caribbean samples previously tested encompass only a portion of the sequence variation seen within the UK Afro-Caribbean population.

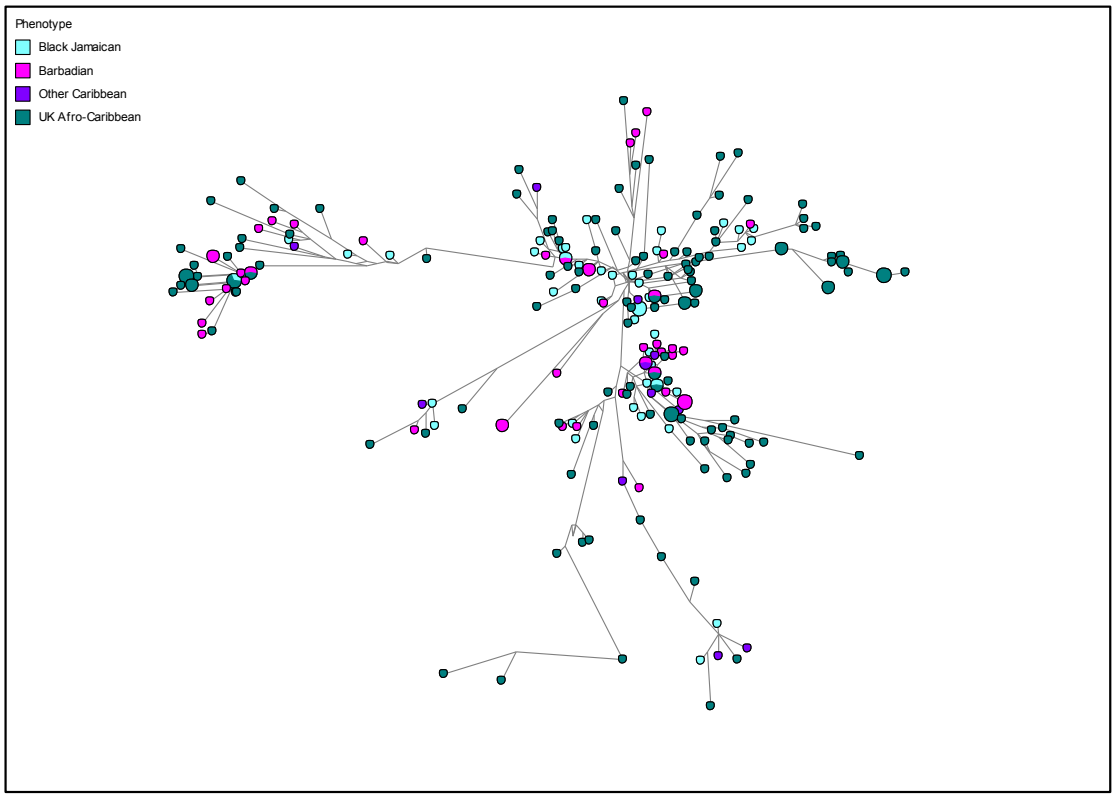


Figure 4.7 Addition of 135 UK Afro-Caribbean mitochondrial sequences to the phylogenetic tree shown in Figure 4.3.
UK Afro-Caribbean samples were sequenced for a minimum of HVI and HVII with just over 40% sequenced for the entire control region. Genotypes for 11 mitochondrial SNPs (defining haplogroups L0-3, N, M, R, H, H1, HV and U) were also used in the construction of this phylogenetic tree, genotypes determined either experimentally for those samples where haplogroup assignment wasn't possible from the initially obtained sequence or inferred for samples where secure haplogroup designation was possible.

Analysing the UK Afro-Caribbean sequences alone produces the phylogenetic tree shown in Figure 4.8. The tree produced contains many discreet clusters that are well separated from each other, demonstrating not just the large sequence variation within haplogroup L, but also the diverse mitochondrial origins of the UK Afro-Caribbean population. There are 6 samples out of 135 that do not classify within haplogroup L. Figure 4.9 provides a closer view of the L3 grouping from which these six non-L haplogroup samples originate (in the top right corner). There is one sequence in close proximity to these non-L samples that's been designated as an L3f on the basis of the control region changes present, however it is well apart from the other L3f samples towards the bottom of Figure 4.9. Additional coding region sequencing of this sample shows that it does indeed have changes at bases 4218, 4350 and 5601 designating it as a genuine L3f sequence (L3f1a in actuality) however all the other L3f samples are much further defined within the L3f lineage (i.e. they have many more change within

the control region associated with branches within the L3f1b grouping), hence placing the two groups apart on the basis of just the control region sequences.

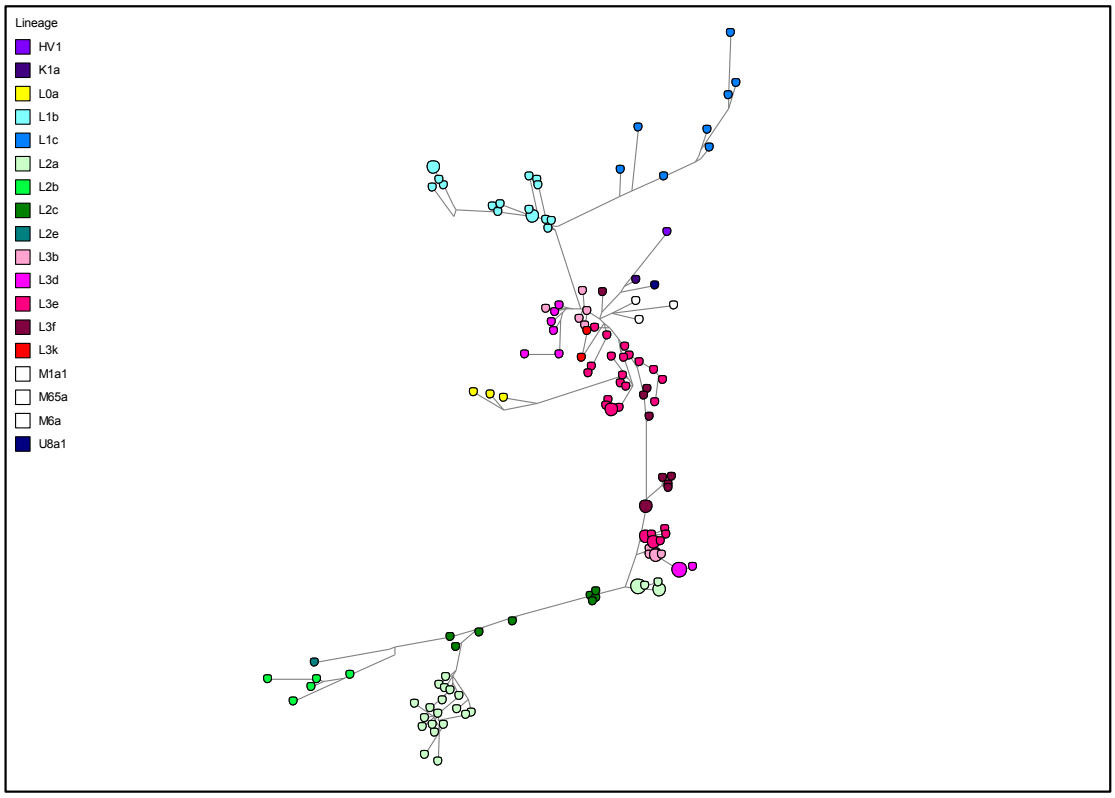


Figure 4.8 Phylogenetic tree of 135 UK Afro-Caribbean mitochondrial sequences
Mitochondrial sequence and SNP data (as Figure 4.7) for 135 UK Afro-Caribbean samples. Dots represent each individual sequence, with the size of the circle proportional to the number of individuals with the same sequence (here the vast majority represent single samples). The sequences are joined together on the basis of similarity in sequence in the most parsimonious phylogenetic tree. The range of colours represents the major haplogroups present. The tree is orientated differently from the previous figures (e.g. haplogroup L1 is now positioned at the top rather than the left hand side).

Of the 6 non-L sequences, 3 (the HV1, K1a and U8a1 sequences) are more commonly associated with European populations suggesting some European admixture within these individuals in their maternal ancestry. Haplogroup U8 is one of the oldest European lineages dating back about 50,000 years while haplogroup K is also a subclade of U8, although this is believed to have originally emerged in the Near East about 30,000 years ago[211]. U8a itself is quite a rare lineage across Europe, with some of the most ancient sequences having a Basque origin, but was found to be totally absent from an analysis of nearly 1,500 North-Africa samples (the only part of Africa with a significant number of non-L sequences)[212]. Of the other 3 non-L sequences, all belonged to haplogroup M, one sitting within sub-clade M1 and the other two in M6 and M65. Haplogroup M is generally found at a high frequency in

the sub-continental region of Asia encompassing, for example, India and Bangladesh, however an exception to this is the M1 subtype which has a very specific distribution encompassing Northeast Africa and the Near East [145, 213]; hence the UK Afro-Caribbean sample with the M1a1 sequence may have Northeast African origins rather than the more common sub-Saharan African ancestry found in people forced into the slave trade, and therefore this individual (or their ancestors) is more likely to have emigrated directly to Britain from Northeast Africa (e.g. Ethiopia) rather than *via* the Caribbean.

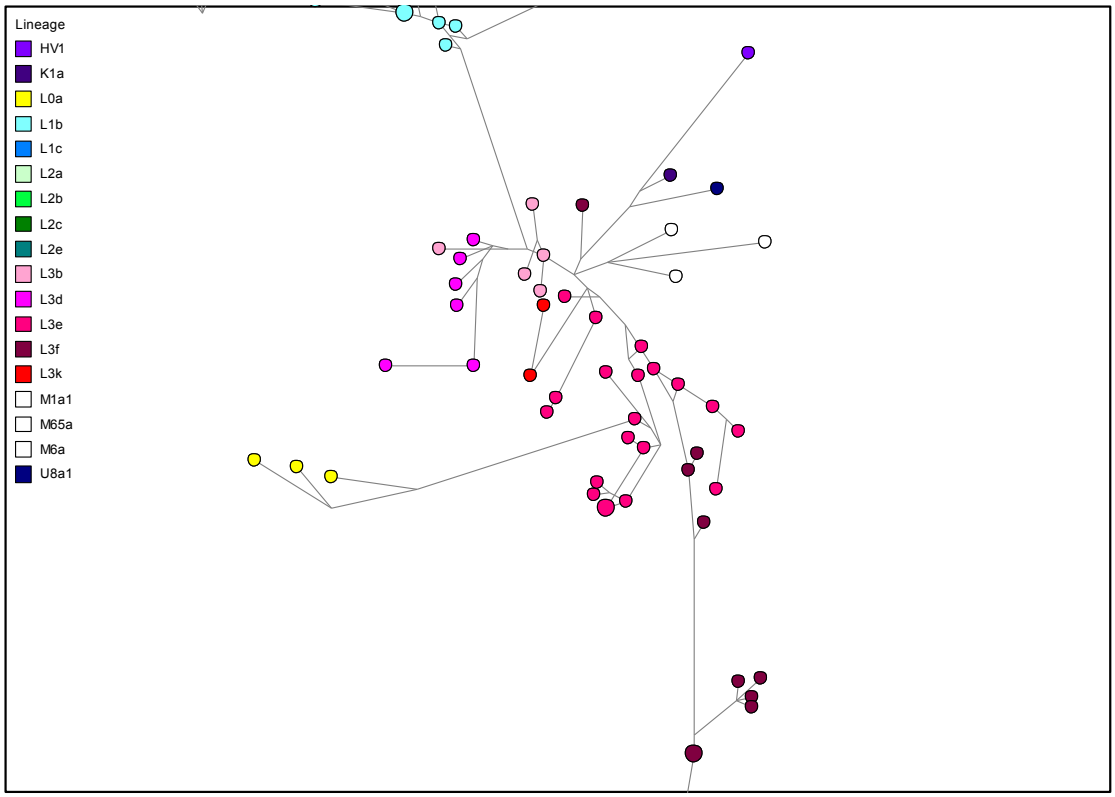


Figure 4.9 Magnified view of the L3 grouping in Figure 4.8.

Figure 4.10 strips away some of the complexity with the individual sequences and instead gives a more basic breakdown of assigned haplogroup based on the control region sequences (and any coding region SNPs if necessary). The high frequency of L3 sequences within the Jamaican population is visually apparent, as is the marked under-representation at L1b. Haplogroup L1b is concentrated along sections of the west coast of Africa from where the majority of slaves were taken for the Atlantic slave trade and is frequently observed in African Americans [52, 214], hence the reduced frequency in Jamaica might suggest that the slaves surviving to contribute to

the Jamaican gene pool preferentially came from tribes along the African coast with reduced L1b levels, in contrast to those enslaved and shipped to Barbados and America. Haplogroups L2a1i and L4b together make up 16% of the Barbadian population (at frequencies of 9% and 7% respectively) yet are not observed at all in the other 191 samples tested from throughout the Caribbean and in the UK Afro-Caribbean population, most likely suggesting a founder effect in the Barbadian population. Whilst multiple UK Afro-Caribbean clusters are shown in Figure 4.7 to be separate from the Jamaican, Barbadian and other assorted Caribbean sequences, this isn't reflected in the more coarse haplogroup breakdown where the UK Afro-Caribbean samples tend to have a fairly similar distribution to the Jamaican samples with a few exception such as haplogroup L1b*.

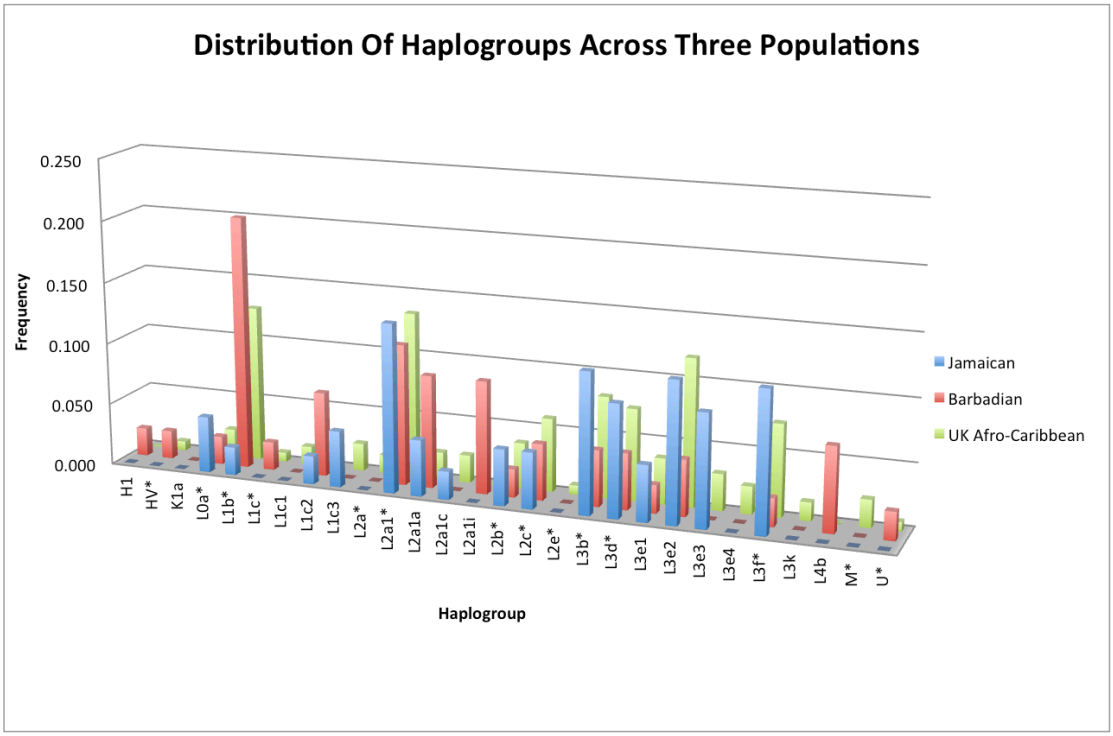


Figure 4.10 Distribution of mitochondrial haplogroups present in the Jamaican, Barbadian and UK Afro-Caribbean samples

Mitochondrial DNA analysis of African American populations has shown only a minor female European influence. In the study of South Carolina African American populations, the African L macro-haplogroup predominated, with 45-55% of individuals belonging to L1 or L2. In contrast the European H haplogroup was observed in only 6 out of 714 individuals [215]. This correlates well with our UK

Afro-Caribbean samples, where 48% of sequences belonged to L1 or L2 and only 3 out of 135 individuals presented with European derived haplogroups (although there were also a further 2 samples with Asian haplogroups). Concurrent analysis of both mitochondrial DNA and Y chromosome markers in the same population has detected significantly different levels of European ancestry when looking at these sex specific DNA molecules. In a wide-ranging genetic study on African Americans, European ancestry was detected at a level of about 28% for the Y chromosome but only 8.5% for the mitochondrial DNA [216]. These findings agree with previous studies [210, 217] and suggest that males and females played different roles in the generation of admixed African –American populations, with the European component coming disproportionately from European men rather than women. Having only examined the more rapidly mutating STRs (rather than SNPs) on the Y chromosome for our UK Afro-Caribbean samples, it's not possible to be sure how well the misclassification rate with the Y-STR method described in section 3.5 reflects the level of European admixture, but with only 11% of these samples misclassifying as European it seems likely that any admixture level is well below this 28% figure quoted for the African-American population.

4.3 UK and Irish Caucasian Populations

The UK Caucasian population, in the main, demonstrates much less mitochondrial variation between individuals than the UK Afro-Caribbean population (where a wide variety of sequences are present). This is visually apparent when comparing the phylogenetic network obtained from the Caucasian individuals (Figure 4.11) with that obtained from the Afro-Caribbean population (Figure 4.8): in Figure 4.8 there is much greater distance (and hence more sequence changes) between clusters of samples.

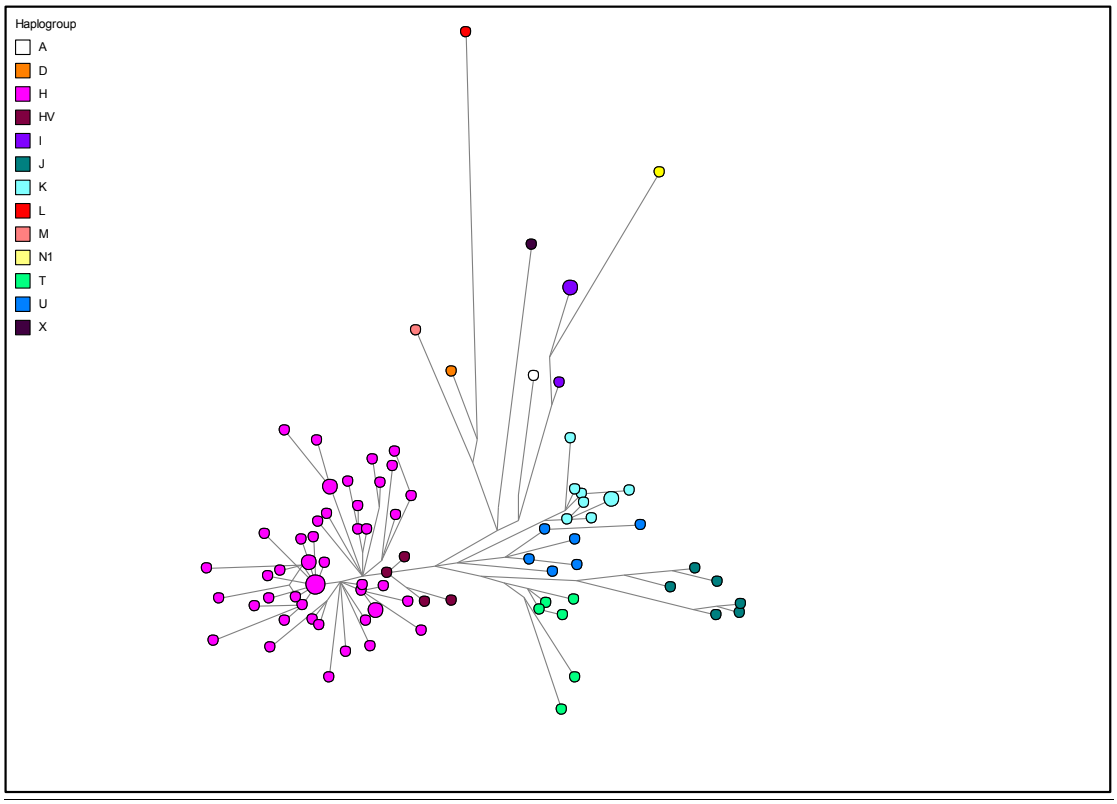


Figure 4.11 Phylogenetic tree of 89 UK Caucasian mitochondrial sequences
Dots represent the individual sequences of samples while the distance between the circles is proportional to the difference in sequence between the samples – enlarged dots indicate more than one sample with the same sequence. This is the most parsimonious phylogenetic tree based on a minimum of HVI and HVII sequencing and select SNP typing (see Table 4.1 for sequences and regions typed).

Haplogroup assignments (and hence sequences) for the UK Caucasian and UK Afro-Caribbean populations show completely different distributions. Approximately 56% of the UK Caucasian samples fall within haplogroup H (totally absent in the Afro-Caribbean samples), all clustering together in a star-like distribution towards the bottom left of Figure 4.11. Many of these samples have minimal, or no, differences between then with respect to the HVI and HVII sequences, and hence further SNP typing was routinely necessary to provide finer haplogroup assignment and aid discrimination. The distribution of H subtypes is shown in Figure 4.12.

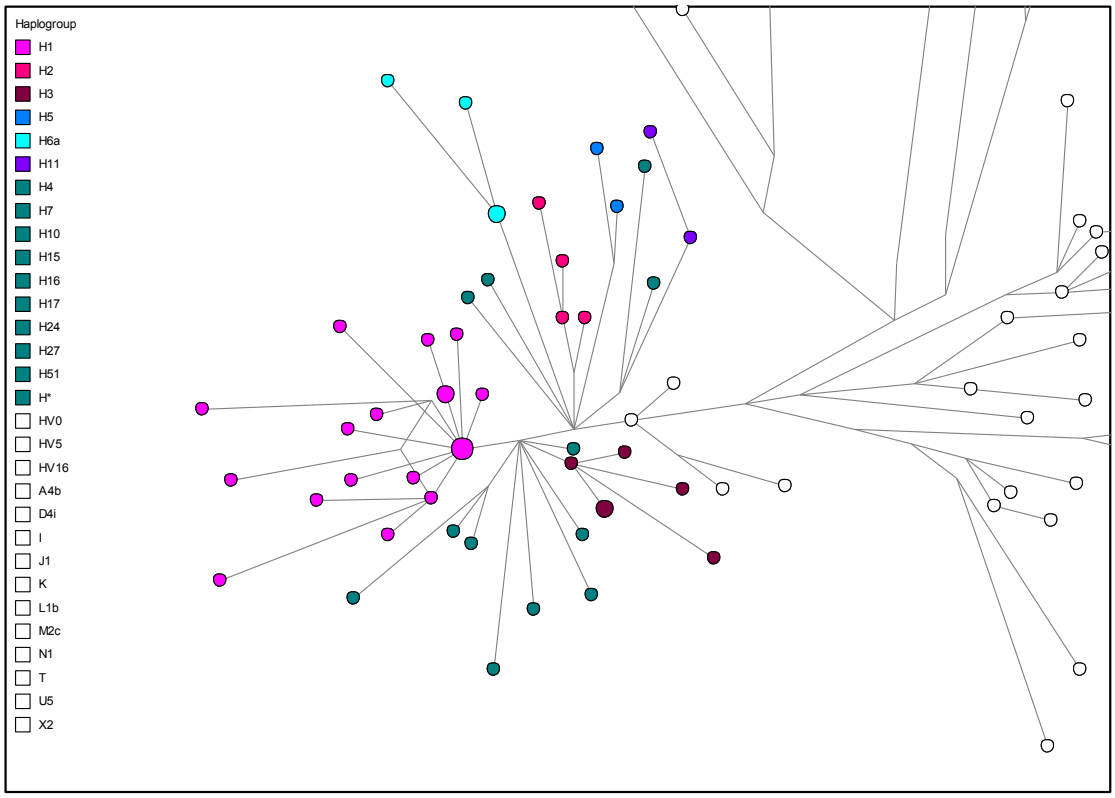


Figure 4.12 Magnified view of the H grouping in Figure 4.11
Haplogroup assignment is now shown in finer detail with all the H subtypes differentiated. All those subtypes represented by just a single individual are coloured green to simplify the diagram. Additionally 3 different samples coloured green are simply designated as H* as these did not map onto any known H subtype despite sequencing the entire 16,569 bases of the mitochondrial genome in these samples.

Originally 8 SNPs were chosen to resolve the H samples into subtypes H1, H2, H3, H4, H5, H6, H7, and H13. These H subtypes were chosen following the publication of one of the earliest extensive sets of full mitochondrial sequences for haplogroup H individuals [149]. Subsequent changes to the phylogenetic tree means that these SNPs now actually assign samples to H1, H2a, H3, H4, H5a, H6a/H17, H7, and H13a1. This SNP typing exercise (along with the control region data) was able to resolve the H subtype of all bar 9 UK Caucasian samples; for these samples the entire mitochondrial genome was sequenced, however 3 of these samples still did not map onto the known phylogenetic tree and hence represent new, previously unknown, mitochondrial variation within the human population. Rather than assigning these novel H subtypes, they are just labelled as H* here. The SNPs tested were chosen from an Italian population dataset [149], and the results presented above show that it would be more useful to substitute the H13 SNP for an H11 SNP in future work (and when using these to increase the discrimination in UK forensic casework) since no

individuals possessed mitochondrial DNA fitting onto the H13 branch of the haplogroup tree, but two samples did map to the H11 branch.

While most samples fell within the common European haplogroups (H, HV, J, K, T and U), there were a few outliers with more unusual designations (seen towards the top of Figure 4.11). Three samples had haplogroup I sequences: this is a haplogroup found at relatively high frequency in Danish Iron Age and Viking Age populations [218], but at a low frequency in other ancient European populations (including Britain [33]), possibly representing a mild Viking genetic legacy although unfortunately the sample size is too small to draw any firm conclusions. Similarly the rare N1a1a mitochondrial DNA type discovered in a medieval Danish cemetery [219] is virtually identical to the slightly more evolved N1a1a3 sequence present in one of the UK Caucasian samples and clustering as an outlier with the I haplogroup samples in Figure 4.11.

The L1b and M2c sequences are most likely the result of unknown admixture at some point contributing African and Asian (respectively) genetic material to individuals of otherwise predominantly White European descent. Haplogroup X2c is generally found at a low frequency (in a meta-study of nearly 22,000 sequences from 66 Eurasian and North African populations it was observed 8 times), and more specifically mainly in Scandinavia, East Europe and West/Central Asia [220], so the presence of a single sequence here is somewhat unexpected. The same is true of the A and D haplogroup samples, with haplogroups A and D being two of the five founding Native American types, having previously left a genetic imprint in Siberia on the journey to the Americas [221]. The specific D4i haplogroup has only been seen before, to my knowledge, in the Yukaghir tribe of the Lower Kolyma and Indigirka delta in Siberia at a frequency of 4.9% [222], in the Fergana region of Uzbekistan at a frequency of 1.9% [223] and at a frequency of 4.8% in the Central Yakutia region of Siberia [224], making it's presence in this UK Caucasian population fairly mysterious. A similar story exists for the A4b sequence where a few individuals have been found to harbour this type in Uzbekistan [223] and Siberia [224, 225], but not in Western Europe.

Analysis of human remains from ancient British burial sites can help put into context the representation of haplogroup U5 within the UK Caucasian population (all 6 samples falling within haplogroup U in the UK Caucasian population sit within the U5 branch). Haplogroup U5a1* was found at a frequency of 19.4% in early (fourth to seventh century A.D.) skeletons, but was absent from individuals found in a later (9th to 11th century) Saxon cemetery, while haplogroup U5b was discovered at a frequency of 3% in the early burial sites and 6% in the later one [226]. This genetic signature is still visible today, as while nearly 60% of the UK Caucasian samples are assigned to haplogroup H, 7% of samples fall within haplogroup U5. Two of these are U5b samples, which correlates to a frequency of just over 2% in a modern population. In keeping with the reduction in U5a1 individuals in the latter half of the first millennium, although 4 samples in the present study fit within the U5a branch, only 1 (1%) is a U5a1 sequence while the other 3 present with variations placing them on independently arising U5a sub-branches and hence cannot be descended from the U5a1 individuals living in early Roman/Saxon Britain.

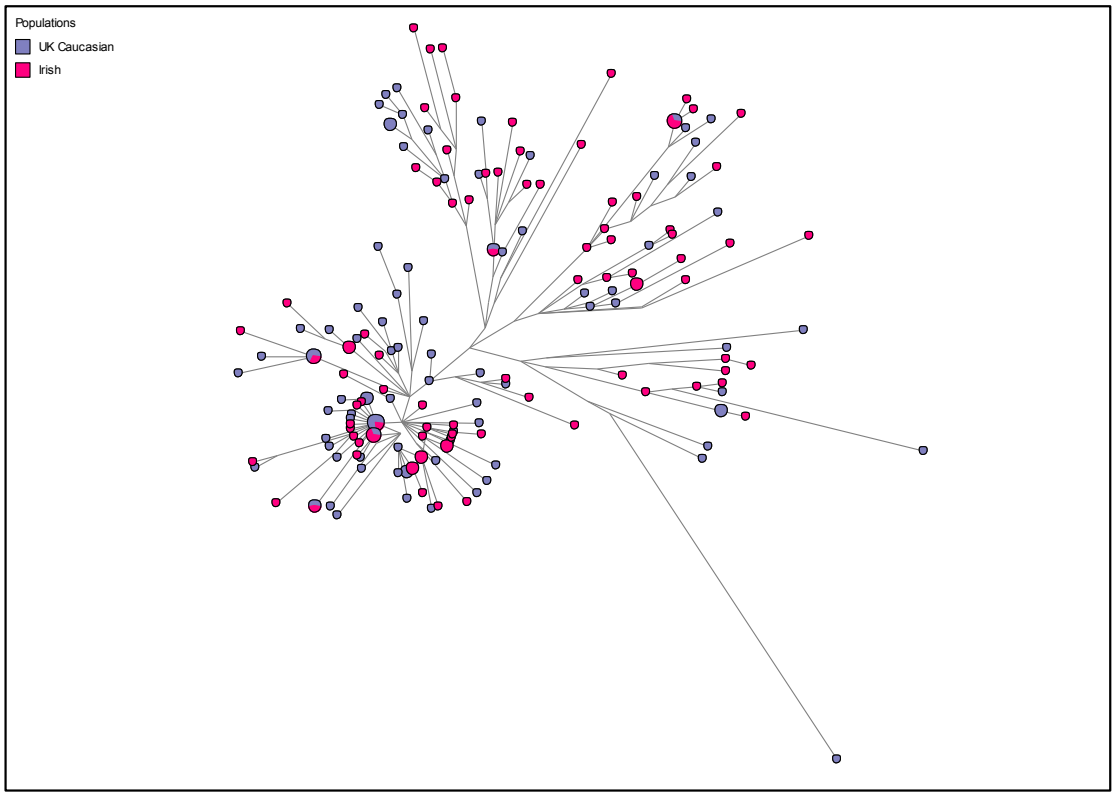


Figure 4.13 Most parsimonious phylogenetic tree of 89 UK Caucasian and 90 Irish Caucasian mitochondrial sequences labelled by population

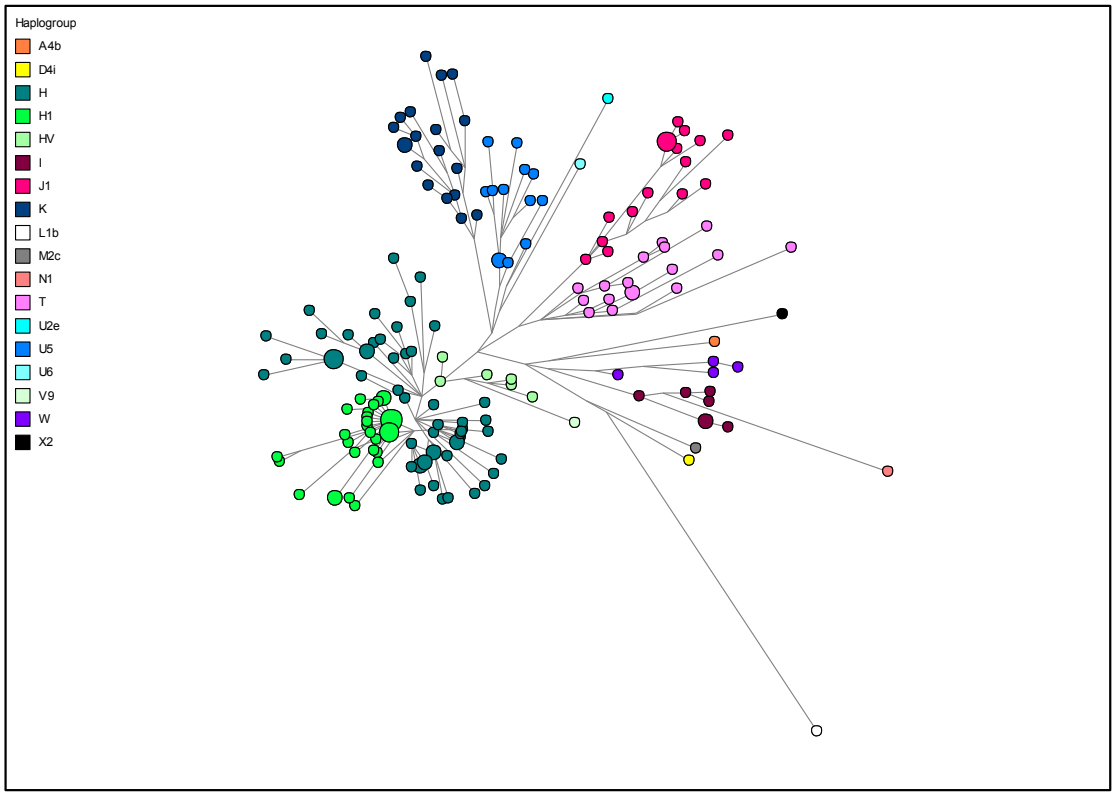


Figure 4.14 Most parsimonious phylogenetic tree of 89 UK Caucasian and 90 Irish Caucasian mitochondrial sequences labelled by haplogroup

Figure 4.13 shows that the mitochondrial sequences obtained from the Irish and UK Caucasian individuals are similar with most sequences branching from a recent common ancestor. There are still some population specific clusters within the groups, for example the separation in the K branch to the top left of Figure 4.13. Figure 4.14 shows exactly the same phylogenetic network, but this time labelled by haplogroup, illustrating that most sequences from both populations fall into the major European haplogroups of H/HV, U/K and J/T with all other haplogroups represented only by those samples to the right of the tree, and of these the biggest group is the previously discussed haplogroup I. The only other Irish samples falling within this extra set are four haplogroup W sequences (in violet to the right of Figure 4.14), a haplogroup not observed at all in the UK Caucasian samples.

On a Europe-wide level, haplogroup W is only observed at a frequency of about 1% [227] compared to the four Irish sequences which represent nearly 5% of the Irish samples. Due to the rarity of this haplogroup generally not a lot of detail is known

about it, although it is a sequence seen at a higher frequency in some populations further East, e.g. specific communities in Iran and Pakistan [228].

Previous research has noted that over 98% of European individuals are derived from haplogroup N (specifically H, I, J, N1b, T, U, V, W and X) [229]. This fits well with our data where 94% of the UK Caucasian individuals fall into this classification while all 100% of the Irish Caucasian samples do. The graph in Figure 4.15 illustrates that the Irish samples are more restricted to certain haplogroup sets than the UK Caucasians, and coupled with the absence of any non-European influences, this reduced genetic variation is similar to that observed with the Y chromosome data from section 3.3 that also showed more homogeneity within the Irish dataset.

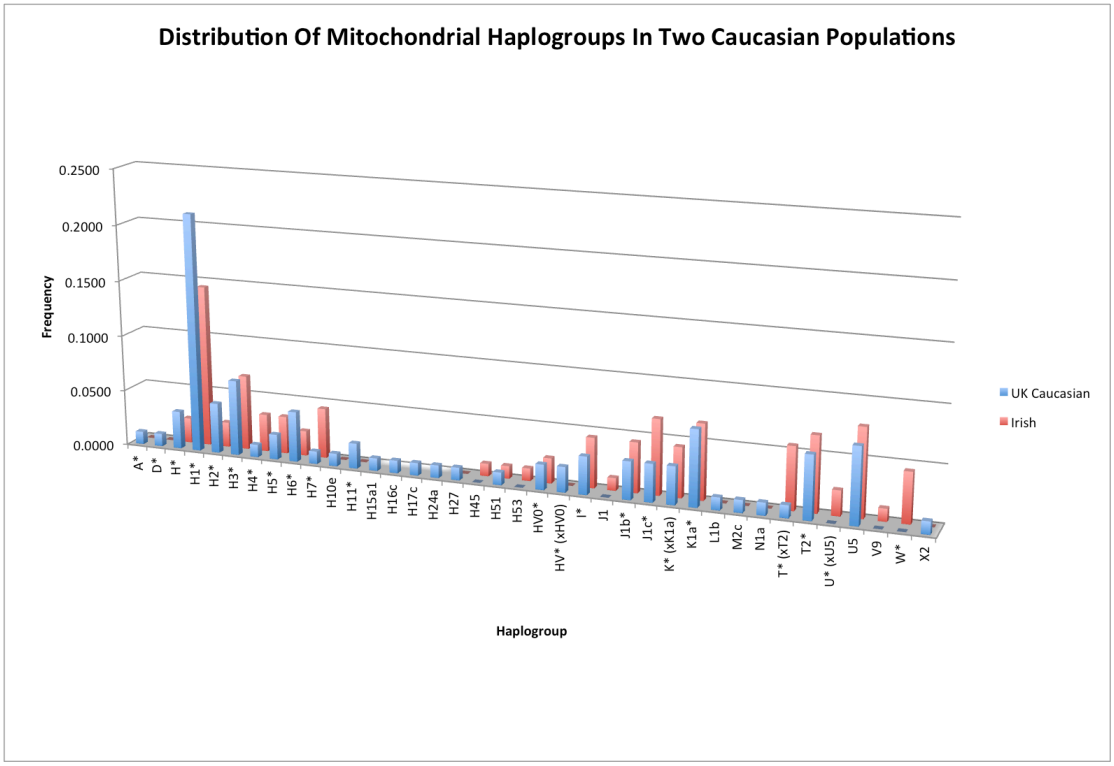


Figure 4.15 Distribution of mitochondrial DNA haplogroups in UK Caucasian and Irish Caucasian populations.

4.4 UK South Asian Population

One hundred and twenty three mitochondrial sequences were obtained from individuals resident within the UK classifying themselves as British Asians. This

generally refers to the region of South Asia encompassing India, Pakistan, Bangladesh and Sri Lanka. The mitochondrial sequences produced classified into a wide range of haplogroups, highlighting the large variation inherent in this group. No sequences were found that fell within the African L haplogroup. Figure 4.16 graphically shows the relationship between the sequences produced, and the haplogroups obtained. The length of the branch is proportional to the number of nucleotide changes between the samples, and it is noticeable how many individual sequences are situated alone at the end of a long branch rather than clustering close together, again showing the diversity within this sample set. This is even the case with many of the M haplogroup sequences in the lower half of Figure 4.16 magnified in Figure 4.17. This is no surprise given the fact that genetic diversity within India is known to be second only to that observed in Africa [230], inline with the early settlement of South Asia following the expansion of modern humans out of Africa.

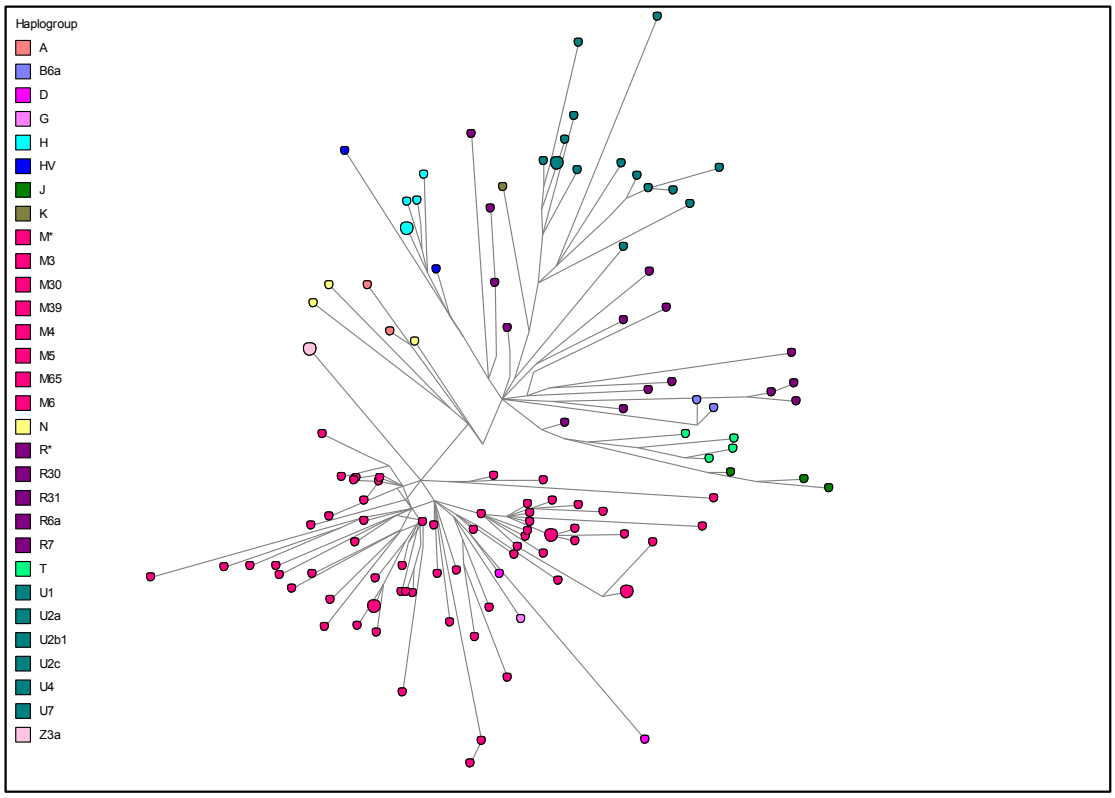


Figure 4.16 Sequence data from 123 British Asian individuals, displayed in the most parsimonious phylogenetic network.
Sequence data is comprised of the entire control region for all samples, plus a small number of branching SNPs that have either been directly sequenced or inferred from the known phylogeny. Full sequence data has not been included for those samples whose entire mitochondrial genome was sequenced as this would have added an extra misleading genetic distance between them and the other samples resulting in their isolation, instead only those SNPs necessary for haplogroup definition have been included.

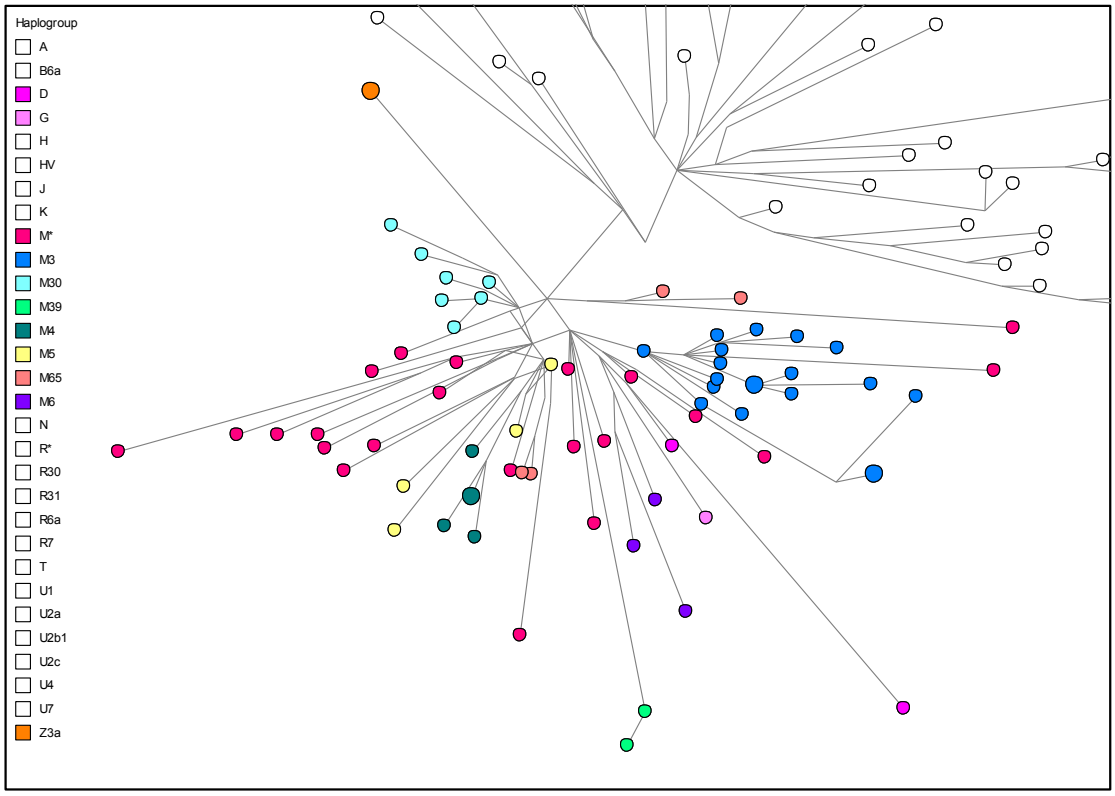


Figure 4.17 Magnification of the lower region of the network presented in Figure 4.16.
This diagram primarily highlights haplogroup M sequences, however this cluster also contains 2 individuals classifying within haplogroup D, 1 within haplogroup G and 2 identical sequences within haplogroup Z3a.

Fifty four percent of all the tested samples fall within macrohaplogroup M. Haplogroup M is observed widely across South Asia [145, 231] with many sub-lineages essentially being restricted to this geographical area (one notable exception being the aforementioned M1 [145, 213]). It has been calculated that of those sub-lineages still present in this part of Asia, the M2 branch is the oldest representing some of the earliest settlers [232], and today comprises 10% of Indian M mitochondrial haplogroups although is only seen twice in this study of UK South Asian samples. Fifteen percent of samples belong to the M3 branch, which is an overrepresentation when compared to a large scale Indian meta-study (of just over 2,500 individuals, M3 was present in 7% of Caste population samples and 7.7% of tribal population samples[231]). In the Indian study the highest concentration of M3 sequences was found in the North and West of India, suggesting that UK individuals of Pakistani descent may be the source of the raised M3 frequency, although without further data this is impossible to verify.

The M3 sequences are shown to the right of Figure 4.17, represented in blue. One undefined M* sequence (shown in cerise) can be seen to cluster with these M3 sequences and branch from a shared point. Given that the mathematical algorithm applied by the network software has grouped this sample within the M3 lineage, this may suggest a previously unseen affiliation not picked-up during the process of haplogroup assignment. The sample (number 434) does indeed possess the basal change from T to C at position 482 diagnostic of M3, and additionally has an unusual change at 16319 which is found deeper within the sub-clade and could suggest it fits within the M3c1b1a branch. Unfortunately the sample is missing a host of other changes defining various branching points leading to M3c1b1a, specifically at positions 152, 16189, 16294, 16179 and 16124 while possessing mutations not known within the M3 phylogeny at 16136, 16217, 16381, 94, 173 and 204. The presence of changes at 16223, 489 and the C to T change observed at SNP 10400 do define this sample as somewhere within the M haplogroup, but without additional sequencing of the sample's full mitochondrial genome it is not currently possible to be more specific about where in the M branch it resides. If it truly does fit within the M3 branch then this suggests there are either unknown reversions within the phylogeny (back-mutations at these bases subsequent to the initial mutation) or that the phylogeny is poorly defined in this part of the tree and needs subtle changes in light of this new sequence. This is a similar situation with many of the M* sequences (and indeed many of the other Asian samples in different haplogroups) where despite the numerous changes observed in the control region with respect to the rCRS, there is no obvious place for the sample to fit on the known M phylogenetic tree. The absence of these observed sequences, or even similar sequences, in the forensic and genetic literature, and hence the gaps in the M phylogeny, point to a paucity of mitochondrial genome sequences from these regions of South Asia.

For a number of the South Asian samples in this project it was necessary to carry out mitochondrial sequencing of the entire genome in order to confidently assign the sample to a haplogroup. This was especially true of some samples where SNP analysis on the pyrosequencer placed them somewhere in macro-haplogroup R but not in the branches H, J/T or U/K. Full sequencing was able to classify them into haplogroups HV, B or specific branches of R. Highlighting the previously discussed

absence of comprehensive sequencing data for individuals from sub-continental Asia, many of these samples will create novel sub-branches on the worldwide mitochondrial phylogenetic tree once published and submitted to EMPOP (the worldwide forensic mitochondrial database) for sequence validation. Some of these newly discovered mutations are established enough in our tested population that it has been possible to see multiple examples of the same unknown coding region changes, for example the two B6a samples both have the same previously undiscovered mutation, in this case since both samples were collected in the Whitechapel area this is most likely to reflect a set of maternal lineages originating in Bangladesh, while all three R6a* sequences also share at least 1 unknown mutation and two of the samples share 6. The R30a lineage currently possesses no sub-branches, and yet both R30a sequences typed in this study have multiple, and divergent, base changes from the known R30a branch, demonstrating that there is abundant diversity within this clade waiting to be discovered.

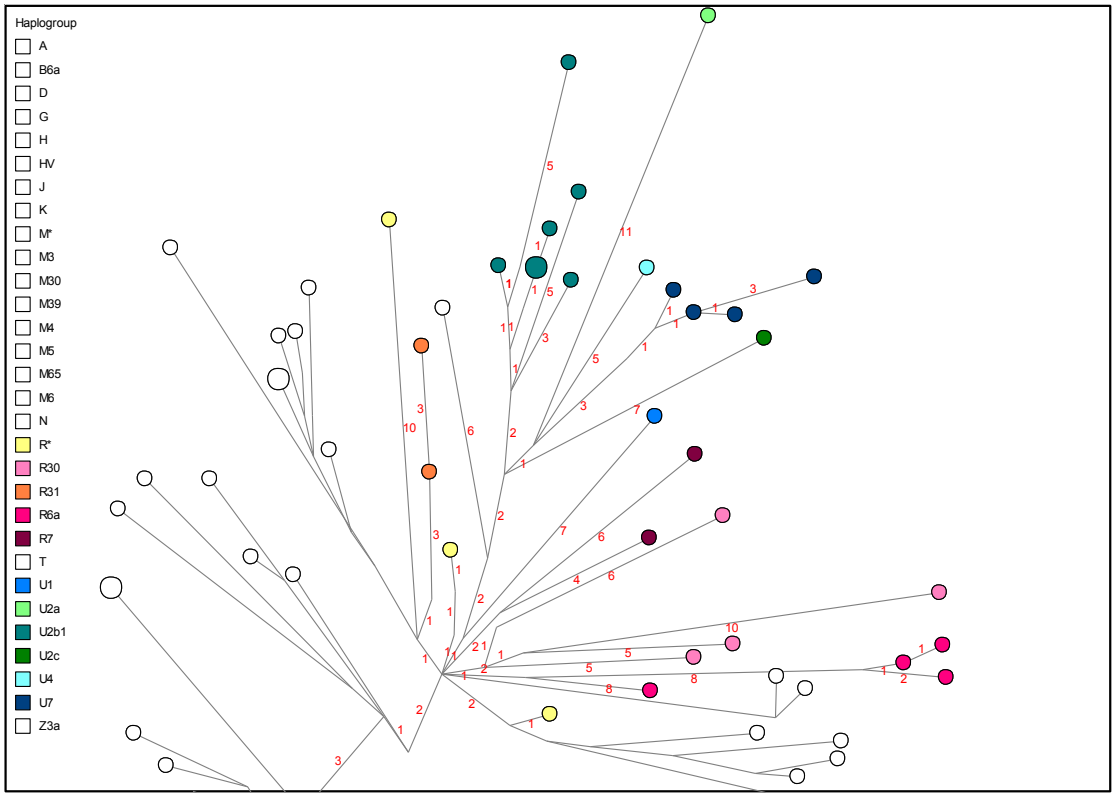


Figure 4.18 Detailed view of the upper section of Figure 4.16 highlighting the sequences classifying within haplogroups R and U.
Numbers in red indicate the number of sequence differences represented by the line, e.g. there are 11 control region mutations between the U2a sequence and the branching point it shares with the U4 and U7 sequences.

Figure 4.18 presents a magnified view of the R and U sequence clusters. The number of mutational steps between sequences is shown in red, and highlights the wide range of R sequences present (the only really similar samples on the basis of control region data are three of the R6a sequences to the right of the figure). Addition of full coding region data would increase these distances substantially, e.g. the R8 sequence, which is shown in the middle of Figure 4.18 coloured yellow, would present with a further 13 mutational changes from the branch point; this is in contrast to the full sequence data obtained for some of the Caucasian samples which on average added an extra 3 changes. Haplogroups R7, R8, R30 and R31 were only discovered in 2004 in a large scale study of the Indian population [147], and here make up 7% of the South Asian samples and 60% of the R types. R2, R5 and R6 complete the R subtypes found and are also known to be present in India from an earlier study [231].

Haplogroups U2a-c are essentially geographically restricted to Southern Asia, being commonly found in India, Pakistan and Bangladesh [228, 231] with coalescence dates in the region of 45,000 years ago for U2a and U2c and 35,000 years ago for U2b (although the confidence intervals for these dates are quite wide) [228]. U7 is another lineage within haplogroup U that is found at a low frequency within Europe, but while this subgroup is observed across sub-continental Asia, it is most prevalent to the west in Pakistan and Iran [228]. Seven percent of South Asian individuals sampled throughout the course of this research belonged to haplogroups U2a-c while a further 3% belonged to U7. A group of similar U2b1 sequences comprising 7 individuals can be seen to cluster together at the top of Figure 4.18. All of these share the same coding region motif of changes at 146, 16051 and 16168, and yet 16168 is not known to be a defining change for this branch, indicating that either the existing phylogeny is incorrect or these 7 samples all belong to a previously unseen sub-branch (there are currently no known sub-branches off U2b1). Figure 4.18 additionally shows that only 2 of these U2b1 sequences match exactly (the enlarged circle) with the rest having various additional changes, suggesting that there is also lots of undiscovered variation within the U2b1 clade.

It is known that in the last 10,000 years there's been migration into sub-continental Asia from the West creating additional diversity on top of the established mitochondrial lineages [147, 232]. This would appear to account for the presence in

the UK Asian population of haplotypes more commonly found in Western Eurasia such as H, K, J and T.

4.5 Population Determination

Due to the non-recombining inheritance of mitochondrial DNA, and hence the sequential accumulation of mutations, groups of individuals sharing ancestry can be linked together or split apart on the basis of specific sequence changes. When these sequence changes coincide with population migration events, specific haplogroups can be restricted/enriched in specific geographical areas, on occasion aided by selective pressures [229]. This is readily seen from the results presented above where each of the different population groups has a different distribution of haplogroups. Figure 4.19 and Figure 4.20 display the phylogenetic arrangement of sequences for all three UK populations, and there is generally good separation between the sequences from individuals of differing ethnicities, albeit some of the separation between the Caucasian and South Asian samples radiating from macro-haplogroup R (encompassing haplogroups U, K, T, J, R, HV and H shown towards the top of the figures) can be harder to discern due to the space available to display that number of samples.

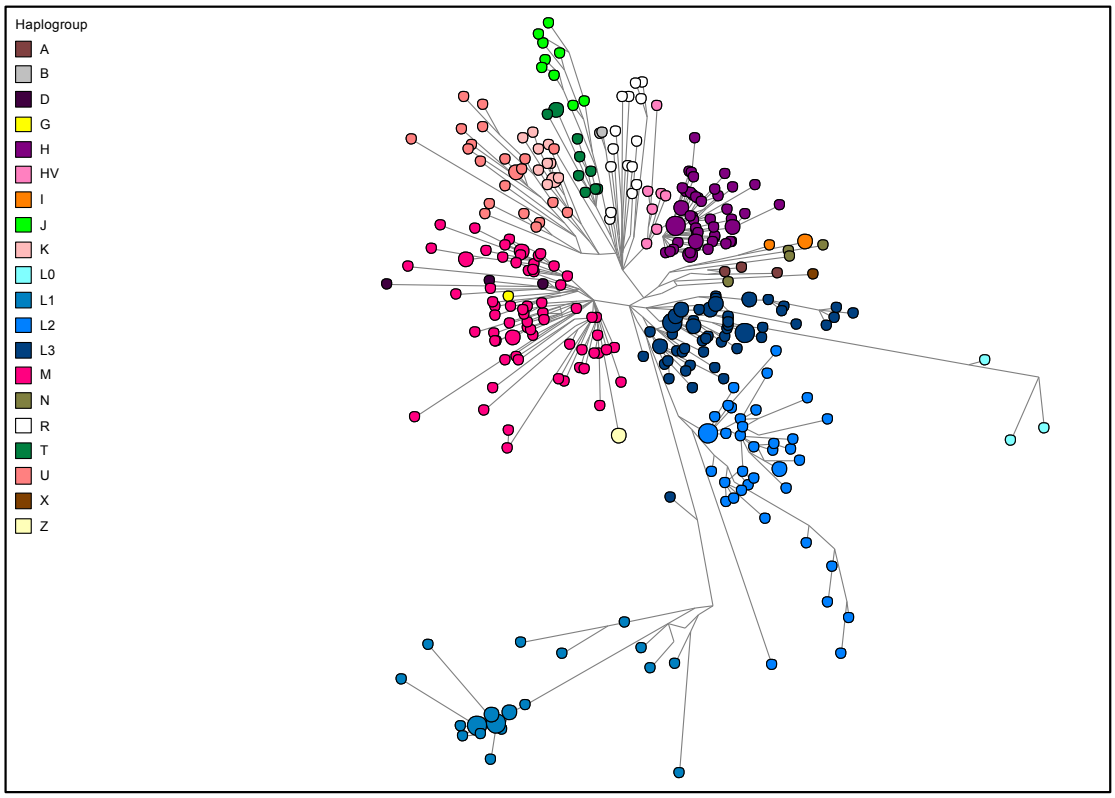


Figure 4.19 Phylogenetic network of 347 UK mitochondrial sequences labelled by haplogroup
In this figure, samples are colour coded by haplogroup. Figure 4.20 is identical but samples are designated by population. The network is calculated based on sequence data from the full mitochondrial control region if available, plus a selection of coding region SNPs either directly typed of inferred from the known haplogroup.

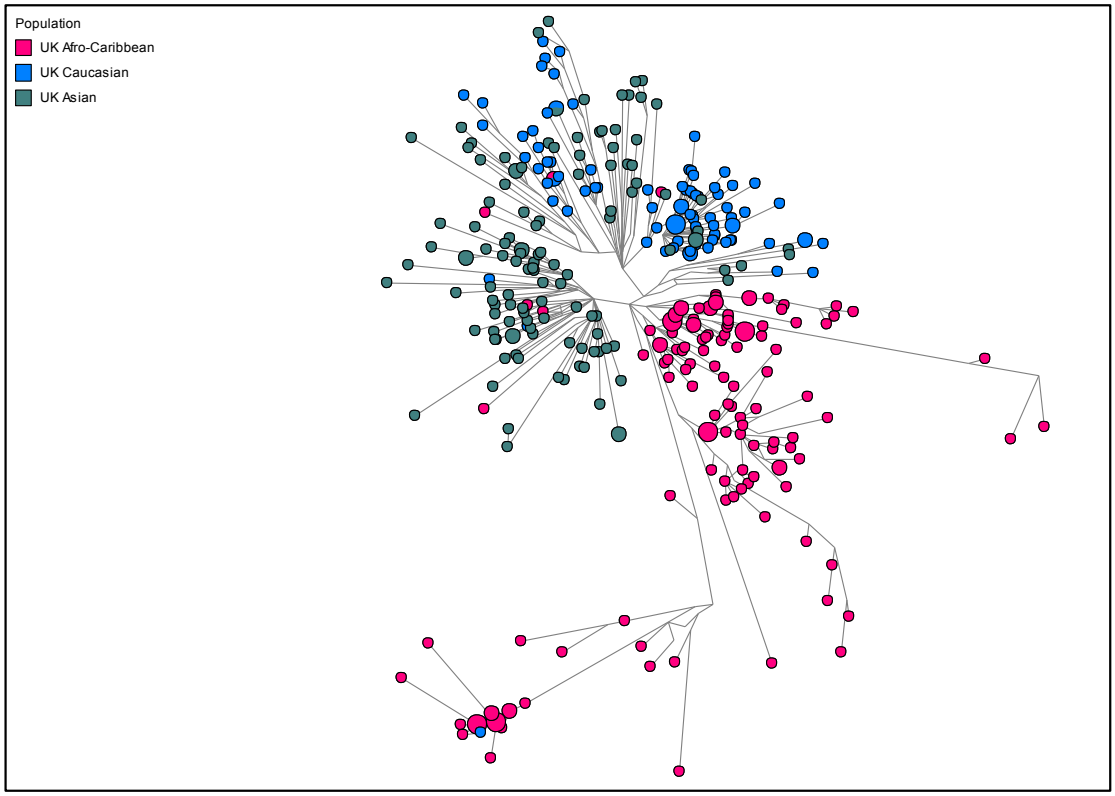


Figure 4.20 Phylogenetic network of 347 UK mitochondrial sequences labelled by ethnicity
Identical to Figure 4.19, except sequences are colour coded by population rather than haplogroup.

Assignment of geographical origin based on mitochondrial haplogroup type is therefore a valid option, and Quintana-Murci et al. [228] have suggested how haplogroups should be linked to location. They assign R0 (including H, HV and V), N1, J-T, U-K, I, W and X to West Eurasia; M*, U2a-c, U9, R*, R1-R2, R5-R6, HV2 and N1d to South Asia; A, B, C-Z, D, F, G and N9a to East Eurasia and L to Africa. This is a slightly blunt classification system missing some of the finer sub-haplogroup detail (e.g. the previously noted presence of M1 in the Near East and Northeast Africa rather than South Asia, the prevalence of U7 in South Asia rather than Europe, and the more recent advances in haplogroup R structure creating more South Asian specific R subtypes), however it provides a good filter with which to look at the data. Figure 4.21 breaks down the haplogroup assignment in this way for the 3 UK populations and highlights the separation of ethnic origin into different haplogroup structures. Success rates for classification using this system would be 96% for both the UK Caucasian and UK Afro-Caribbean samples and 74% for the UK South Asian samples.

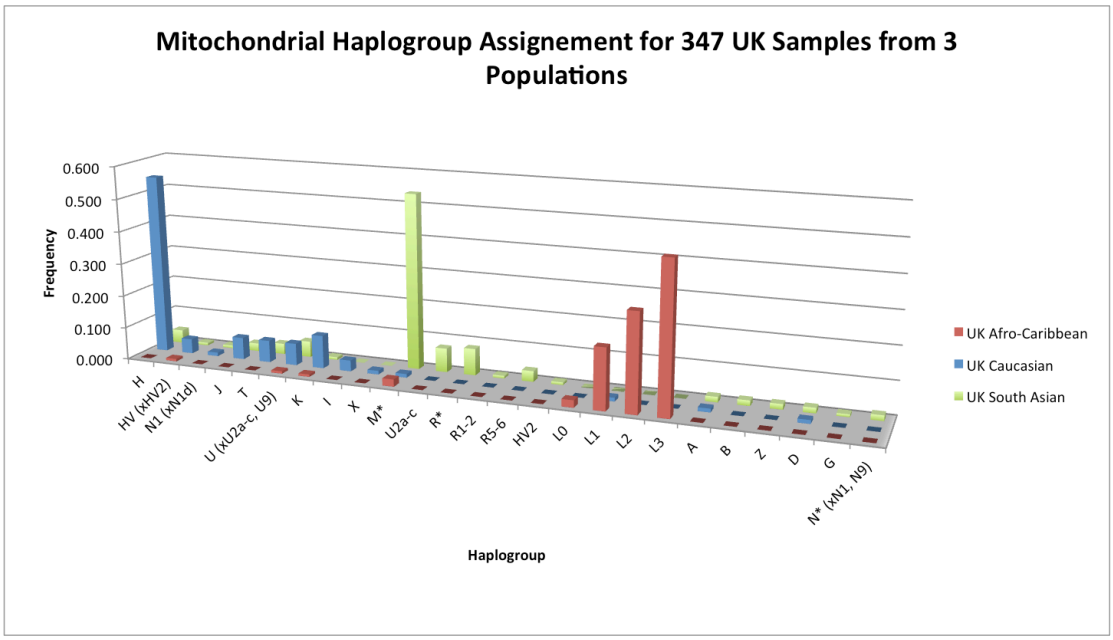


Figure 4.21 Mitochondrial haplogroup designation for 347 UK samples from 3 populations

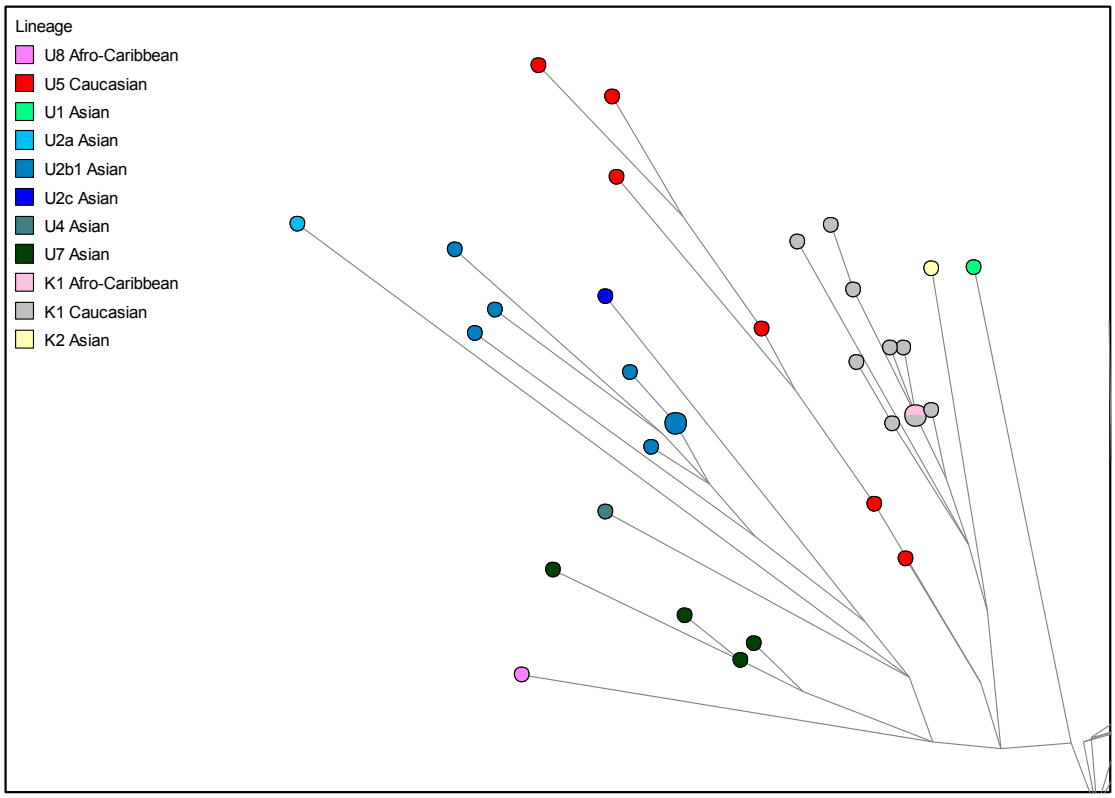


Figure 4.22 Magnification of the U (and K) cluster from Figure 4.20

As mentioned above, a more detailed analysis of haplogroup sub-structure can improve this classification to an extent. Shown in Figure 4.22 is the haplogroup U section from Figure 4.20. There is quite clear delineation here in some branches between populations, with U2a-c and U7 clearly associated with South Asia (respectively seen 9 and 4 times in the South Asian population but never in the other two ethnicities) and U5 associated with the UK Caucasian population. It is not possible from this data to draw any firm conclusions about clades U1, U4 and U8, but the literature would suggest a fairly complex distribution for U1 while the European roots of U8 have previously been discussed in section 4.2 hence a conservative adjustment to the classification system would just be to designate U7 lineages as more likely to be South Asian than African or Caucasian in line with previously published work [228]. Additionally designating M1 individuals as more likely being of African descent rather than Caucasian or South Asian is also in line with conventional wisdom, while further sub-defining R lineages as South Asian or East Asian would be necessary were we interested in a 4 population classification rather than working with only the 3 most common ethnicities present in modern British (and hence for this purpose excluding individuals from East Asia, principally China). Table 4.3details the

results from this updated classification system, demonstrating a high success rate for correct ethnic origin determination in the UK when analysing an unknown DNA sample originating from one of the three main populations present within the country.

Table 4.3 Population of origin classification success using mitochondrial DNA haplogroups

Actual	Predicted				
	Caucasian	Afro-Caribbean	South Asian	East Asian	Unknown
UK Caucasian	96%	1%	1%	2%	
UK Afro-Caribbean	2%	96%	1%		
UK South Asian	14%		77%	7%	2%

Further work could refine this prediction system by examining the sub-haplogroup structure in more depth to ascertain whether more precise boundaries could be drawn with respect to haplogroup-population affiliation (this is especially applicable to the South Asian samples, e.g. two of the three South Asian J samples sit within a specific J2 branch, while all 17 UK or Irish Caucasian samples belong to J1), or by designating samples that fall within some haplogroups as 'not-determined' if the prediction is likely to be uncertain. Any such refinement would require the additional sequencing of an independent set of samples from each UK ethnicity to validate any classification decisions based on this data rather than rooted in previously published studies. Due to the nature of uniparental markers however, there is always going to be a certain error rate due to admixture that it's impossible to account for (going back 10 generations the information obtained from a uniparental marker is only representative of one ancestor out of about a thousand that has contributed to the genetic background of that person). Hence one solution is to trade the detailed knowledge of a single lineage (in this case maternal) within a person's ancestry for a broader, if shallower, understanding encompassing the entire spectrum of genetic inheritance from all contributing lineages. One way to achieve this is through the use of autosomal SNPs.

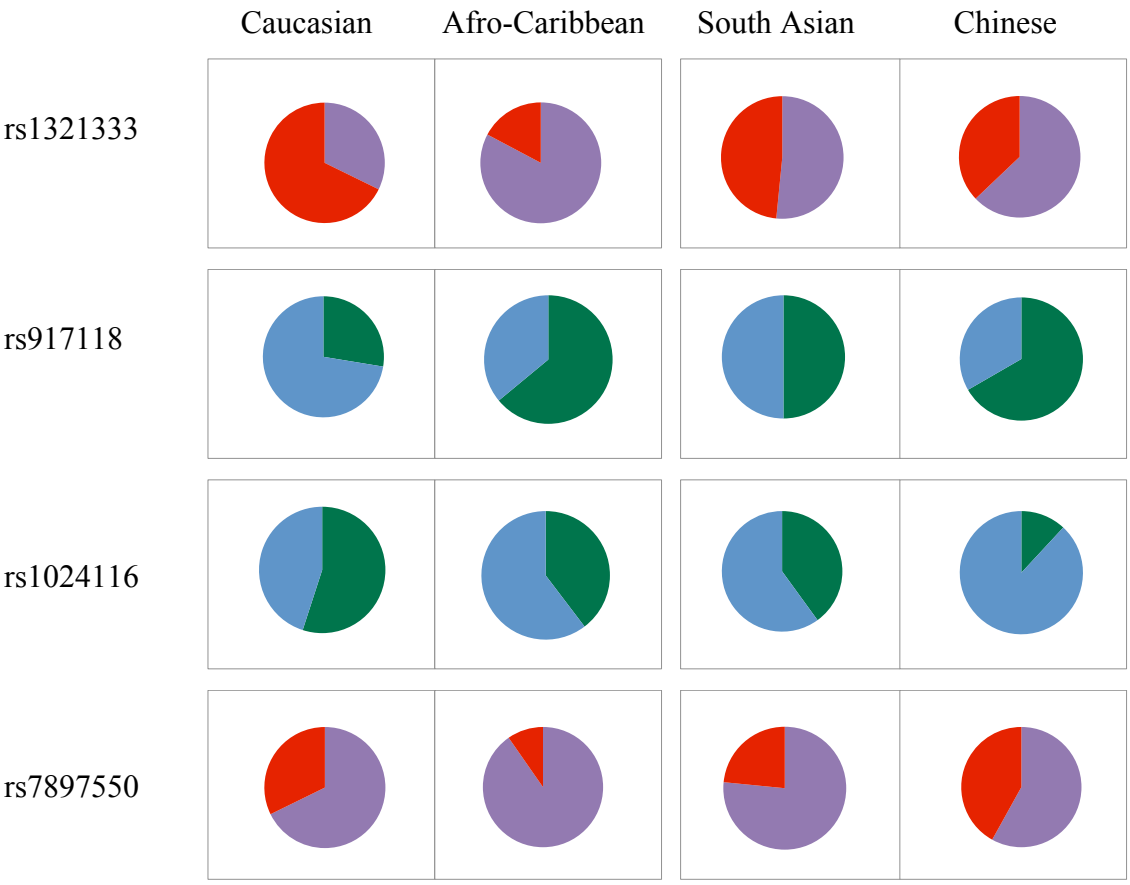
5 Population Specific SNPs Results and Discussion

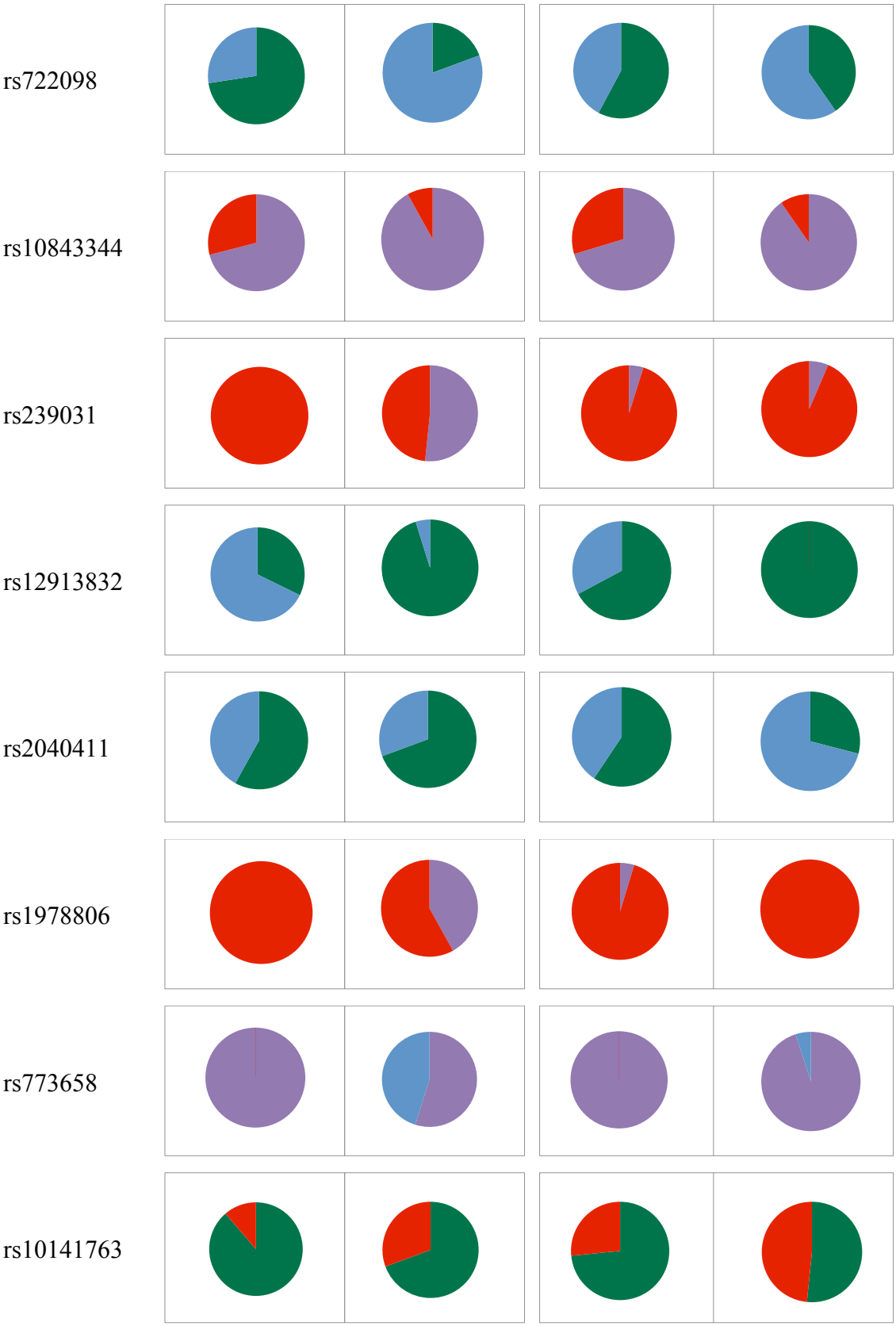
As chapters 3 and 4 have shown, in most cases single lineage markers can be sufficient to obtain a good classification of an individual's self-stated ethnicity, however a better test would utilise information supplied by a more representative cross-section of all those ancestors.

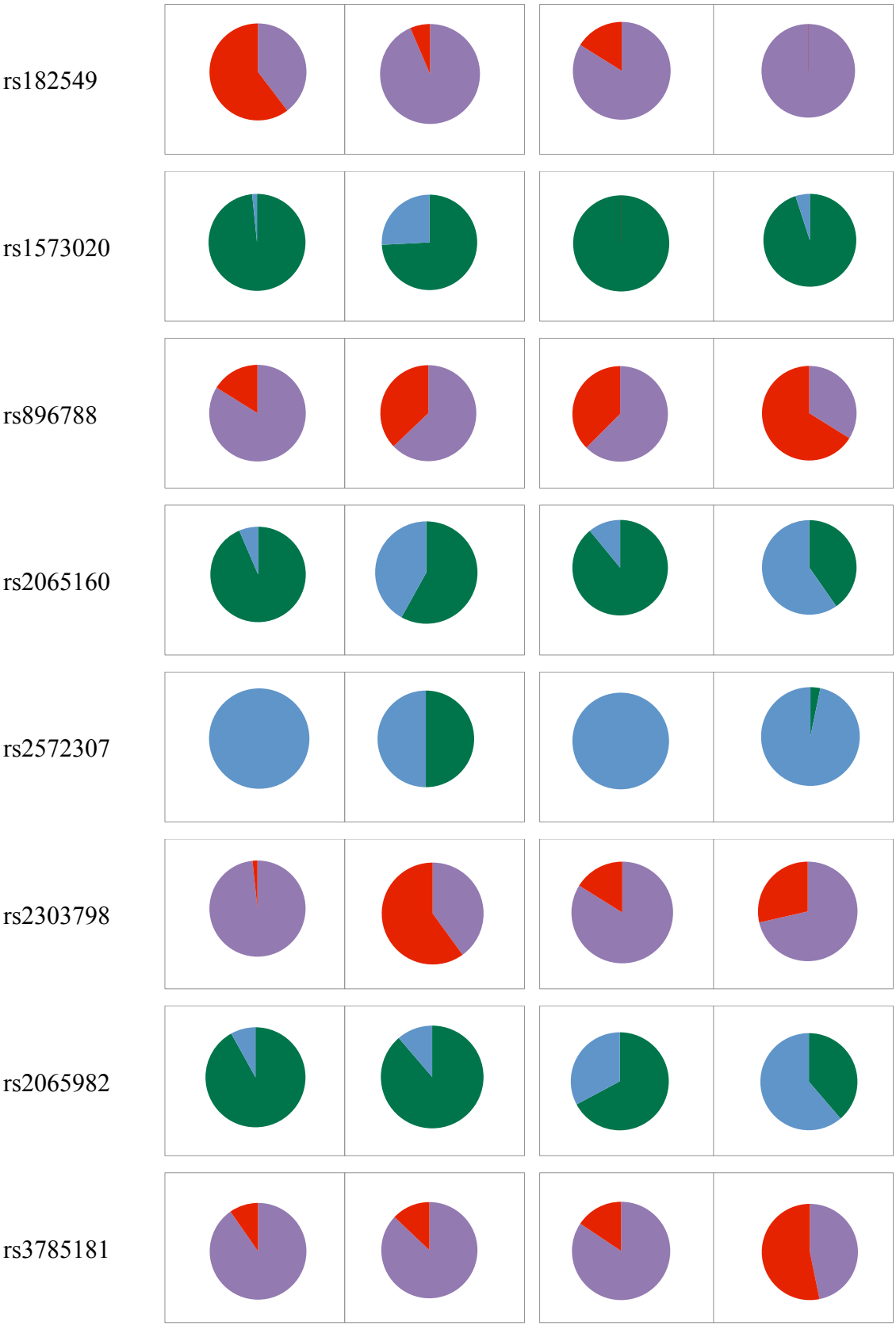
Section 1.3.3 detailed a number of different autosomal genetic markers that could be used for this purpose. While all have been shown capable of differentiating individuals from different continental groups, some of these markers have more practical drawbacks than others, especially were these tests to eventually find use in a forensic setting where DNA quality and quantity is often limited. The high mutation rate of STRs means that allelic differences between populations (e.g. caused by founder effects) will break down more quickly than with more slowly mutating markers, and hence many STRs are needed to provide a robust classification system [43]. From a practical viewpoint this raises issues not just with analysing so much data but also with assay design as the allele range of each marker means there is a finite limit to the number of fluorescently labelled STRs that can be multiplexed together in one reaction, consequently requiring a great many reactions to be performed per sample to get results for something in the region of these 377 STRs analysed by Rosenberg *et al*. [43]. Copy number variations involve the differential presence of large (>1kb) sections of DNA, however most methods for detecting these copy number variants are relatively complex and expensive [18, 233], and even real-time PCR results require normalization [233]. Alu repeats are much easier to genotype, however unlike SNPs, there is not the same wealth of information available to aid marker selection for those loci demonstrating skewed allelic distributions between different ethnic groups. Hence, due to the ease of genotyping with many different platforms, the abundance of published and freely available data on SNP frequencies in different populations, and the tolerance of many SNP typing platforms to poor quality (degraded) DNA, sets of SNP markers were selected, and assays designed, with the aim of developing a more accurate population classification system than those previously described in chapters 3 and 4.

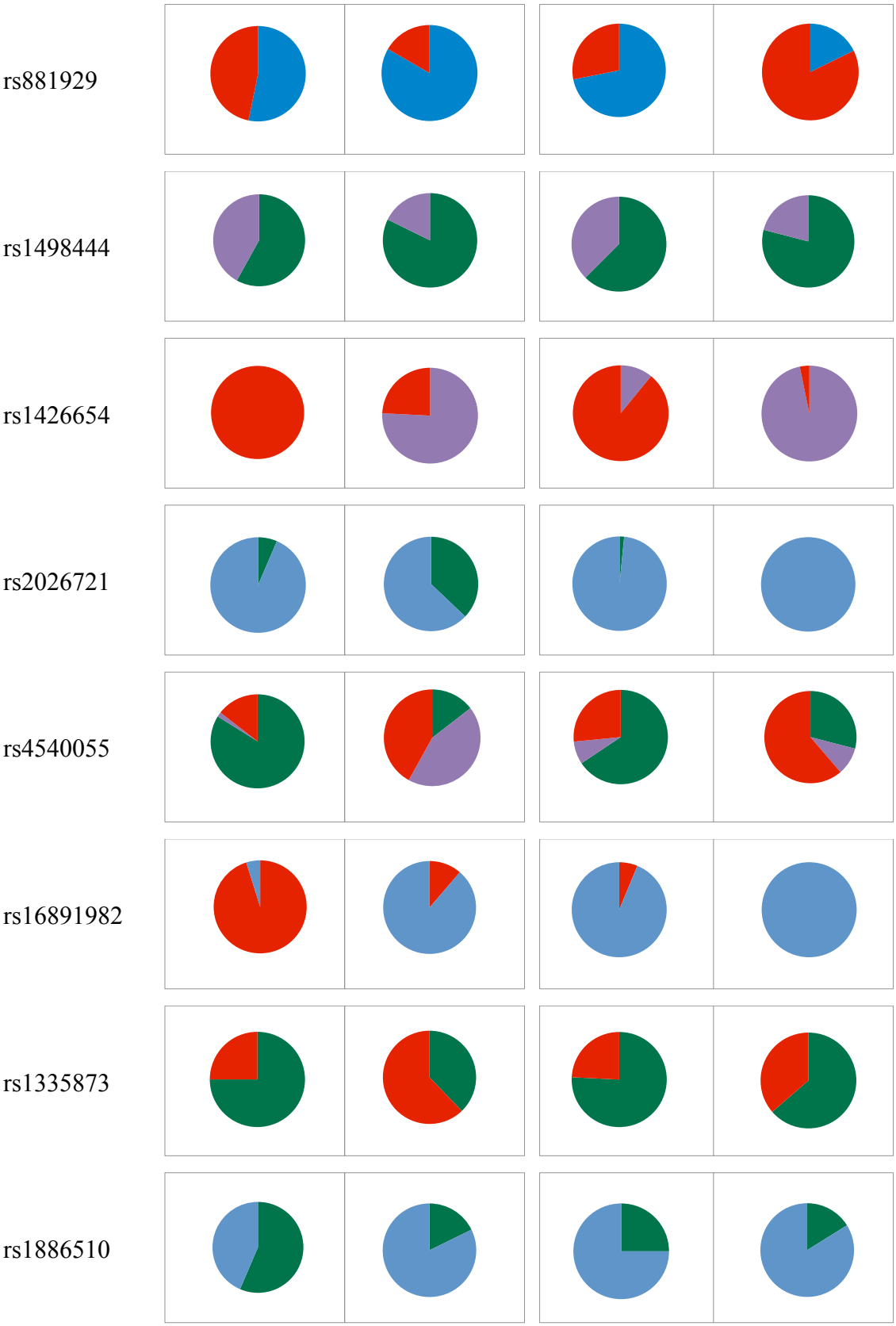
5.1 Individual Markers

An initial set of 22 population specific SNPs (ancestry informative markers) were chosen and a multiplex assay designed as detailed in chapter 2. Genotype results from a wide range of samples were produced in a collaborative effort from laboratories in Santiago de Compostella, Cologne, Copenhagen and London where we specifically tested samples from British residents of Caucasian, Afro-Caribbean, South Asian and Chinese ancestry. Allele frequencies derived from this data differed from those reported on dbSNP (used in candidate locus selection) for some markers. This lead to the removal of some SNPs including rs312895, rs3793599, rs928221, rs2303891, rs13045690, rs1662821, rs34650568, and rs1004788. Additional SNPs were substituted for those removed and following validation of these markers the set increased to first 25 loci, then 32 and finally 34. Allele distributions for these final 34 SNPs across our four British populations are shown below in Figure 5.1.




rs722098
rs10843344
rs239031
rs12913832
rs2040411
rs1978806
rs773658
rs10141763


rs182549
rs1573020
rs896788
rs2065160
rs2572307
rs2303798
rs2065982
rs3785181


rs881929
rs1498444
rs1426654
rs2026721
rs4540055
rs16891982
rs1335873
rs1886510

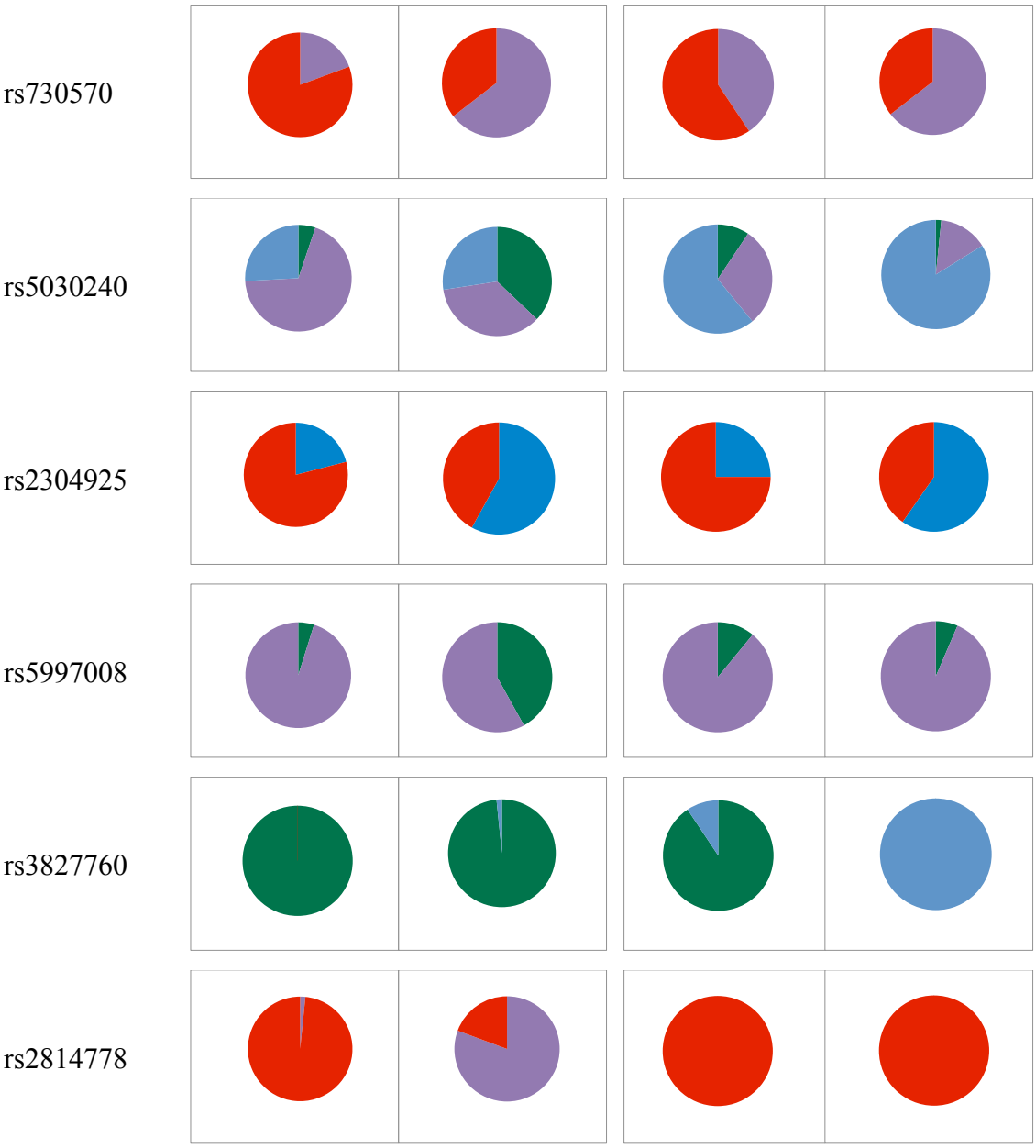


Figure 5.1 SNP frequencies in 4 British populations for 34 markers.
Green represents allele A, purple allele C, blue allele G and red allele T.

There are striking frequency differences between populations for many markers, e.g. for the bottom two SNPs in Figure 5.1 it can be seen that the minor allele is almost exclusively confined to one population; for rs2814778 the C allele is found at a frequency of 81% in Afro-Caribbeans while entirely absent from the Chinese and South Asian populations and only present at a frequency of 2% in the Caucasian population, and for rs3827760 the G allele is absent from the Caucasian population, found at a frequency of 2% in the Afro-Caribbean and 9% in the South Asian populations, yet is the only allele present in the Chinese population. These types of

frequency disparity between populations are utilised below to develop a Bayesian classification system.

Some of the markers are of more limited value when tested with these populations, for example rs1498444, although even this marker has a variance in the minor allele frequency between 0.18 in the Afro-Caribbean population and 0.42 in the Caucasian population.

Following analysis of the data and allele frequency generation, a chi-squared test was used to check that the genotypes were in Hardy-Weinberg equilibrium (i.e. that there was the correct distribution between homozygotes and heterozygotes) for each marker in every population. The majority of markers were shown to conform to the Hardy-Weinberg principle in all populations, although there were a few instances where the chi-squared test was significant: most of these involved very skewed distributions where the presence of a single opposing homozygote was the anomaly (e.g. for rs182549 in the Afro-Caribbean population the genotypes were CC=28, CT=2 and TT=1). In all of these cases the offending genotype was rechecked for accuracy. There was one more significant departure from Hardy-Weinberg equilibrium, and that was in the South Asian population for rs12913832 (AA=18, AG=7, GG=7). All the genotypes were rechecked for this marker, but once again the analysis was found to be correct, suggesting that we may well be seeing a signature of stratification within this population (which is not unexpected considering the large geographical area from which the individual's ancestry may be drawn).

As expected, all SNPs presented with a maximum of two alleles in any one population except for the two tri-allelic SNPs (rs4540055, rs5030240) where all three known alleles were observed in the course of this study. These tri-allelic SNPs provide quite good population differentiation, and place fifth and seventh respectively if all SNPs are ranked by their divergence estimates (i.e. how useful they are) – a list of Jensen-Shannon divergence estimates [234] is generated automatically by the Snipper app when analysing a group of training sets. The four most useful SNPs are shown to be rs3827760, rs1426654, rs16891982, and rs2814778 and over half of the total divergence obtainable with this 34-plex marker set (across these four populations) is

generated by the 8 most useful SNPs while the 10 least useful SNPs only increase the divergence by 9%.

5.2 Three Population Classification

The SNPs in this marker set were originally chosen to maximise divergence between Western European, African and East Asian populations. Using individuals resident in Britain with either Caucasian, Afro-Caribbean or Chinese origin, training sets were produced each containing 31 samples (62 alleles) to provide meaningful minor allele frequencies for these predominantly di-allelic markers.

The data was run through *Structure* [42] as detailed in section 2.9.3 which employs a Bayesian clustering method to discern genetic structure within groups of samples. The results of this analysis are shown in Figure 5.2. Each vertical line represents one sample, and the colour composition of that line reflects the proportions of each calculated genetic cluster (i.e. highlighting any admixture of the sample). The number of hypothetical genetic clusters that structure searches for is user defined, and this is the *K* number. Figure 5.2 shows results for K=2, 3 and 4. At K=2, the Caucasian population is clearly distinguished from the other two, indicating that this is the most genetically dissimilar group from the rest on the basis of these specific 34 SNPs. The *Structure* program defines the genetic clusters without prior knowledge of population affiliation, hence the fact that the Caucasian population has been successfully separated verifies that the chosen SNPs provide good Caucasian:Non-Caucasian differentiation (the samples are grouped together by self-declared ethnicity in the diagram for simplicity, with the three populations separated by black lines).

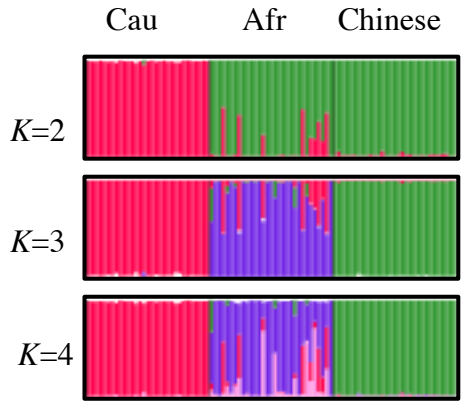


Figure 5.2 Structure plots for the 34-plex SNP set in British Caucasian, Afro-Caribbean and Chinese populations
Each sample is represented by a vertical line, and the colour composition of that line denotes the calculated membership to each of the Bayesian derived cluster. The *K* parameter denotes the number of hypothetical genetic clusters the model was attempting to discern from the data. Samples are grouped by self-declared ancestry, left-to-right being Caucasian, Afro-Caribbean and Chinese. More detail can be found in section 2.9.3.

At *K*=3 (i.e. the expected number of populations) the three groups are quite clearly separated, with complete distinction of the Caucasian and Chinese populations. The Afro-Caribbean population is clearly defined too, with all individuals associating at least 50% with the third (purple) *K* cluster and most associating much more strongly than that. It is clear however that the Afro-Caribbean population is showing less homogeneity, with a number of samples demonstrating low-level contributions from other populations, mainly here the red Caucasian cluster. If *K* is increased to 4 (i.e. searching for extra unknown genetic clusters additional to the three pre-defined ethnicities) then it becomes clear that further structure can be detected in the Afro-Caribbean populations, and the admixture in these individuals is no longer predominantly with the red (Caucasian defining) cluster, but instead between clusters 3 (purple) and 4 (pink) which are both highlighting different Afro-Caribbean genetic signatures. If the increase in *K* had not found any genuine additional genetic structure then patterns would be seen similar to those displayed for *K*=5 in Figures 5.4 and 5.5 for the 33-plex and 34-plex results rather than the specific distribution seen above in Figure 5.2. The presence of this additional genetic structure in the Afro-Caribbean population is not surprising given the large variation within Africa (and to a lesser

extent the Caribbean), as previously discussed with respect to the mitochondrial results.

A Bayesian population assignment algorithm similar to that used in *Structure* is implemented in the Snipper App Suite (http://mathgene.usc.es/snipper/), a simple web-based portal for rapid classification of individual profiles into one of a set of pre-defined populations [186]. In a similar way to the non-admixture model in *Structure*, allele frequencies are calculated for each population, and then the queried profile placed into the best-fitting population. Unlike *Structure*, allele frequencies are directly calculated from user-defined training sets and then values are given for the likelihood of the queried profile belonging to each of the specified training sets on the basis of the entire 34-marker genotype. When using the genotypes from the Caucasian, Afro-Caribbean and Chinese data as training sets, the classification 'apparent success' rate is 100%, i.e. when population level allele frequencies are generated for each marker using the data of the 31 samples contained within each training set (in this case they will be identical to those displayed in Figure 5.1), then when those same 93 profiles are assessed in their entirety (the 34 SNP, 62 allele genotype) and entered into the classification algorithm then all profiles classify into the correct self-declared ethnicity. This is not a surprise given the good separation observed in the structure results above in Figure 5.2. If a cross-validation analysis is performed then classification success is also 100% - this involves removing each of the 93 samples in turn from the training sets, recalculating the allele frequencies from the remaining 92 samples, and then classifying the removed sample using the modified model. Hence, this 34-plex SNP system is shown to be able to correctly classify samples belonging to UK individuals of Caucasian, Afro-Caribbean or Chinese ancestry with 100% success. Details of the final 34-plex SNP package, the implementation of the prediction algorithm, and the classification success rate with a three population system have been published [186].

5.3 South Asian Specific SNPs

During the preliminary validation work for this set of population specific SNPs, it became clear that there were significant problems distinguishing between Caucasian

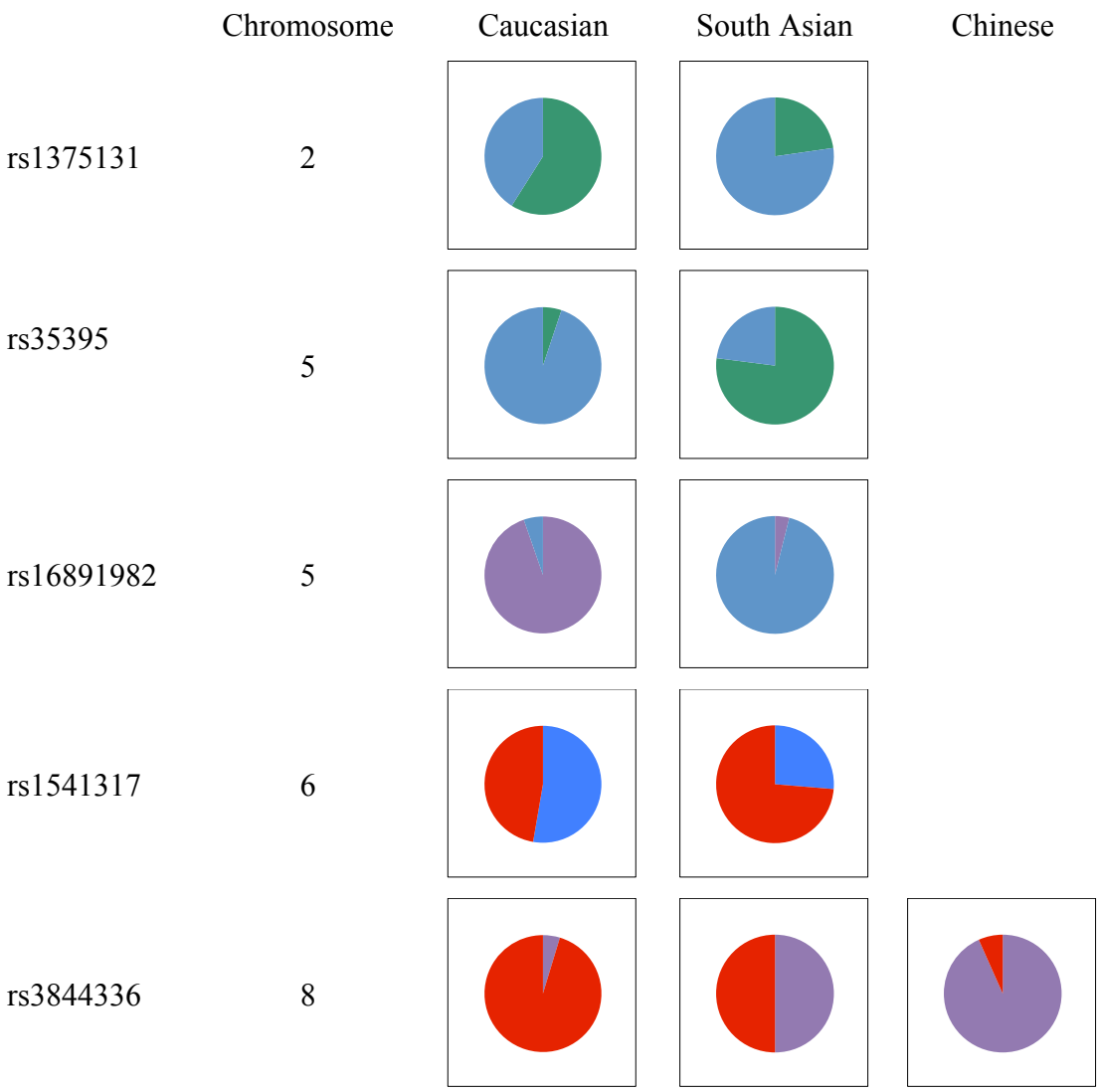
and South Asian individuals; not a specific problem for our other collaborators who rarely test individuals with Asian sub-continental ancestry, but an essential requirement for use of this test within Britain given the population demographics here.

Selecting SNPs with a characteristic South Asian genotype can be challenging. Sub-continental Asia (principally India, Pakistan and Bangladesh) was first colonised by *Homo sapiens* about 70,000 years ago with Sri Lanka following about 20,000 years later [235]. Some migration west from Burma may also have occurred 4,500 to 11,000 years ago [235]. Due to religious, tribal and linguistic differences, geographical and climatic features (for example deserts and mountains), and cultural reason (e.g. the Hindu caste system which stratifies some parts of the population by status) there is a great deal of population substructure within South Asia. This has been shown genetically through the mitochondrial genome where clear differences can be seen in India between members of different castes [236, 237] – interestingly the higher castes show more genetic similarity to Central Asian populations than the lower castes [237]. In contrast, a study of more than 1000 autosomal markers (STRs and indels) in 15 distinct linguistic or cultural Indian subpopulations showed only a modest genetic differentiation between the groups [238]. A comprehensive multicentre study was undertaken to address some of these issues and concluded that there was indeed a high degree of differentiation observed between 55 different Indian subpopulations when studying 405 autosomal SNPs [235]. Hence care has to be taken when choosing and validating South Asian specific markers that the allele frequency differences observed between continental groups are not confined to specific South Asian sub-populations.

In a study by Yang *et al.* [183], 195 ancestry informative markers (AIMs) were tested against a range of different populations, but while 75 of these were shown to be useful in discriminating between European and African individuals, 72 between European and Amerindian individuals and 113 between European and East Asian individuals, only 3 were useful for separating Europeans from South Asians. Additionally, clustering analysis showed that when the South Asian samples were included in a 4 population split (essentially European, African, Amerindian and East Asian) then they presented as an admixed population principally affiliated with the European cluster

with only a minor East Asian contribution. It's therefore expected that separating genetically the British South Asian samples from the British Caucasians may be the most challenging task with a small number of SNPs. Some of the markers showing the highest differentiation between Caucasian and South Asian populations are skin tone markers.

Using pyrosequencing technology, nine promising candidate SNPs were chosen from the publications of Yang *et al.* [183] and Seldin *et al.* [35], along with one from a study by Bauchet *et al.* [182] and additionally a SNP known to be functionally associated with light skin-pigmentation in Europeans (specific information on the choice of these markers is given in section 2.9.1. The frequency results from these markers are shown below in Figure 5.3 and preliminary data published [239].



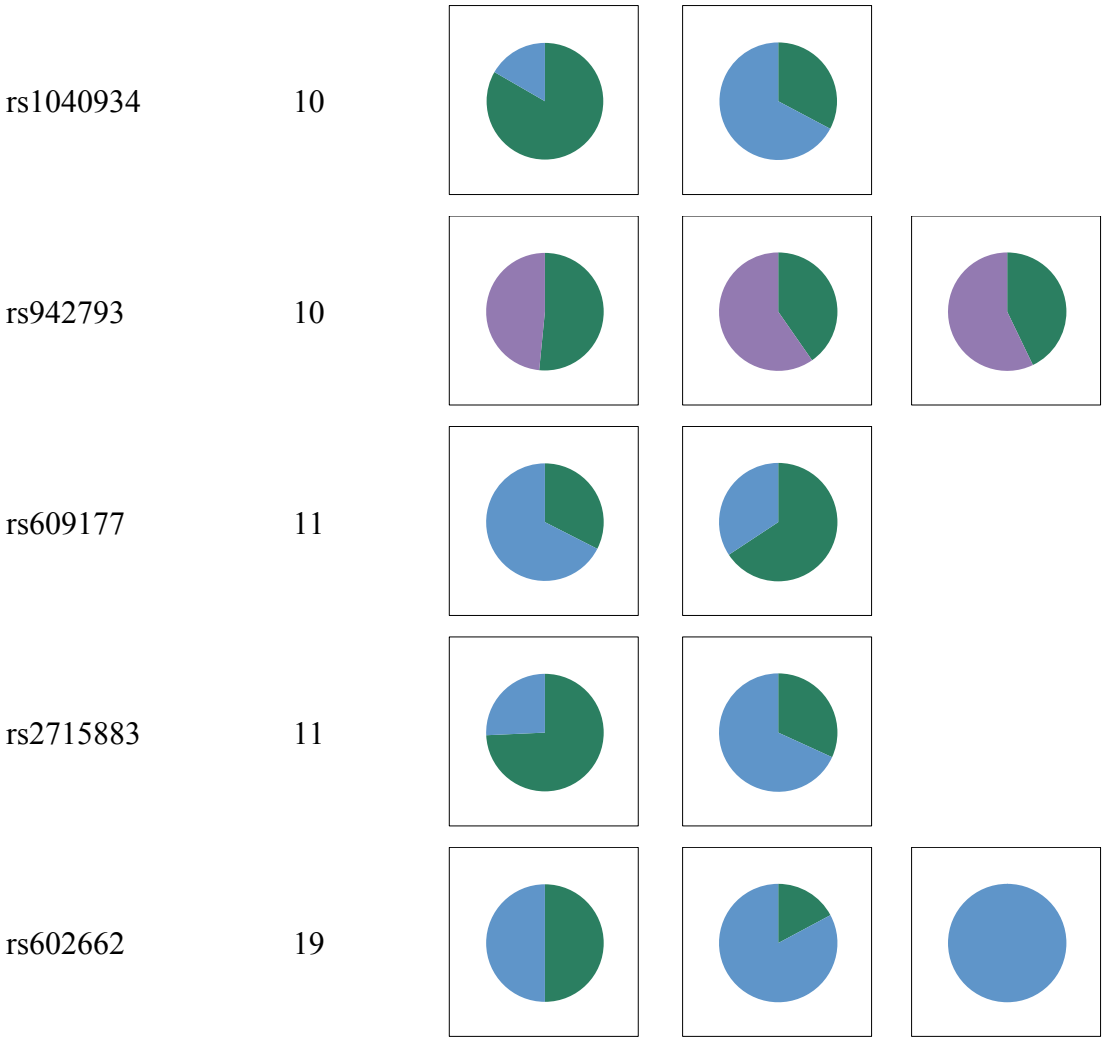


Figure 5.3 Graphical representation of allele frequencies for 11 candidate Asian specific SNPs
Green represents allele A, purple allele C, blue allele G and red allele T.

Primer design proved complex, and genotyping result poor, for rs911903 due to the sequence context of this SNP (designed primers had a high tendency to bind strongly to multiple locations on the genome), and hence no further work was carried out on this marker following the initial set of validation tests. In the Seldin study [35] this SNP showed the fourth highest divergence between European and South Asian individuals.

SNPs rs1375131, rs1541317, rs1040934 and rs609177 are the four SNPs successfully typed from the Seldin study and all show skewed distributions where the major allele in one population is the minor allele in the other population. Of the Yang SNPs, rs3844336 and rs602662 both show good Caucasian-South Asian divergence, and additionally present with an even better differentiation between Caucasian and

Chinese populations. SNP rs2715883 from the Yang list is another marker with a validated British Caucasian-South Asian allele difference ($p<0.001$) and turns out to be located in an intron within the POU transcription factor which is primarily expressed in the epidermis, being especially important in keratinocyte differentiation and proliferation [240] – hence this SNP may be in linkage disequilibrium with a specific allele(s) for this gene that may have a phenotypic impact.

Of those remaining SNPs, rs942793 proved to be of very limited use with A/C allele frequency distributions of 0.52/0.48 for the Caucasian samples and 0.40/0.60 for the South Asian samples. This SNP is supposed to show a variation in allele frequency across Europe between the Northern and South Eastern populations [182], however this evidently doesn't translate to a marked European/South Asian difference. Interestingly when tested on our Chinese samples the allele frequencies were very similar to those obtained from the South Asian samples despite the high European/East Asian divergence observed with the genotypes submitted to dbSNP from the Affymetrix 10k SNP mapping array [241].

Two SNPs (one from the Yang list and the known pigmentation SNP) were co-located within the same SLC45A2 (solute carrier family 45, also called MATP) gene on chromosome 5. This gene codes for a membrane associated transport protein involved in melanin synthesis and is known to be associated with human pigmentation variation [184, 242]. A study of over 3 million polymorphisms from the phase 2 HapMap data release suggested that changes within this gene were under positive selection in the European population [243]. The DNA sequence change effected by SNP rs16891982 causes a non-synonymous coding change in exon 5 of the protein, with an amino acid substitution of phenylalanine with leucine at position 374. The phenylalanine allele is associated with light skin in Europe while the leucine variant appears to be the ancestral allele present at a high frequency in other worldwide populations [185, 242, 243]. This SNP also appears to play some role in skin tone variation within Europe [184, 242]. SNP rs35395 is located in an intron of this gene.

Given the known association of rs16891982 with lighter skin tone, particularly in the evolution of European individuals, and the positive selection that has been exerted on

this gene within European populations, it is perhaps no surprise that the two SNPs within SLC45A2 show the best allele divergence between the Caucasian and South Asian individuals tested here. The difference at rs16891982 is particularly striking with C/G allele ratios of 89:5 in the Caucasian population compared with 4:98 in the South Asian samples. The results achieved with this rs16891982 SNP were so useful, that it was included within the final 34-plex primer set (the results produced across all populations for this marker on a different sample cohort can be seen in Figure 5.1).

The additional use of rs35395 is more problematic due to the proximity to rs16891982 on chromosome 5. The classification models all work on the assumption that the markers used are in linkage equilibrium. Some of the candidate SNPs showing good Caucasian-South Asian differentiation are located on the same chromosome and may potentially be linked. Linkage values can be estimated using the HapMap data, which shows that SNPs rs609177 and rs2715883 on chromosome 11 have a recombination rate (the chance that a recombination event in meiosis will split, i.e. unlink, the two loci) of 0.06, SNPs rs942793 and rs1040934 on chromosome 10 have a similar recombination rate of 0.053 while SNPs rs16891982 and rs35395 on chromosome 5 have a recombination rate of only 0.00076. The rates around 5% would result in linkage between these two markers when looking at close family relatives, however any linkage will quickly break down at a population level and hence all four markers on chromosomes 10 and 11 could be used without adversely affecting the classification system. The rate of 0.076% however will mean that the two SNPs within the SLC45A2 gene will definitely be in linkage disequilibrium within a population and hence only one of these markers can be used in the classification system.

The South Asian samples used in this study mainly originate from India, Pakistan, Bangladesh and Sri Lanka as individuals from these countries are the main contributors to the Asian population within the UK, however the recent increase in immigration from Afghanistan may alter this slightly, and it is likely that Afghan genetics will be slightly different again from the other populations (due to the geographic border, and the largely dominant Muslim religion in both countries, the Pakistan and Afghan populations are expected to show the most similarity). The large heterogeneity of those describing themselves as British Asian also must warrant some

caution when using pigmentation SNPs in a Caucasian/Asian classification since there is a difference between the lighter skin tones more commonly seen in Afghan individuals with the darker skin tones more common in Sri Lankans, and all the gradation between in the Indian, Pakistan and Bangladesh populations. SNP rs16891982 demonstrates just such a change with a statistically significant variation in allele frequency between South Asian individuals presenting with dark and light skin [244], although it should still be noted that this difference reflects a change in the allele C frequency from 97% in the darker skinned cohort to 83% in the lighter skinned cohort while our Caucasian data is still highly differentiated with an allele frequency of only 5%.

5.4 Four Population Structure

Presented below are results for a four-population classification (Caucasian, Afro-Caribbean, South Asian and Chinese) with three different combinations of SNP marker set. The 33-SNP set shows the results obtained prior to the Asian sub-continental SNP research detailed above that resulted in rs16891982 being included within the final 34-plex marker set. The 36-plex marker set consists of the standard 34-plex with the addition of an extra two South Asian SNPs selected from those tested in section 5.3 (rs1040934 and rs2715883).

Structure plots of the resulting data are displayed in Figure 5.4, Figure 5.5 and Figure 5.6.

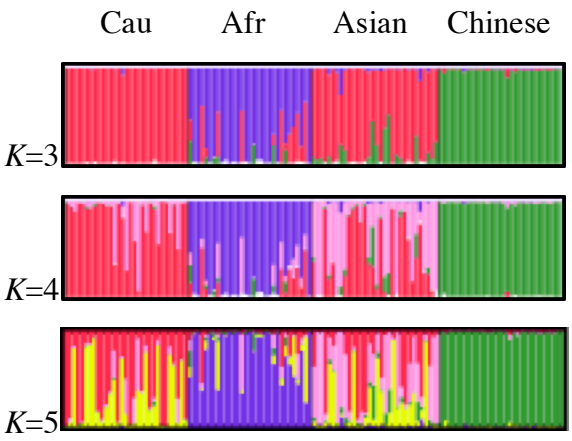


Figure 5.1 *Structure* plot of 33-SNP marker set.
Samples are grouped together by self-declared ancestry – left to right Caucasian, Afro-Caribbean, South Asian, Chinese. The *K* parameter represents the number of theoretic genetic clusters that the model searched for – each cluster represented by a different colour.

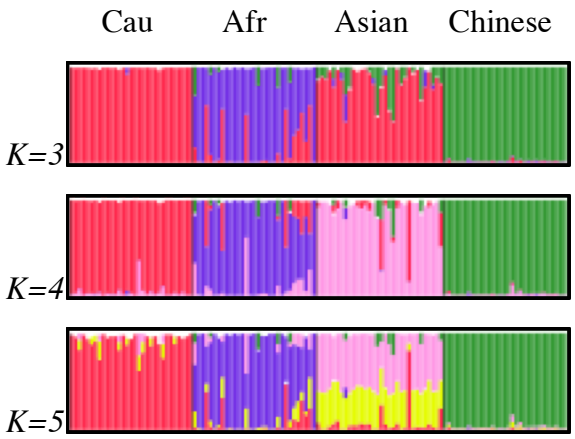


Figure 5.2 *Structure* plot obtained with data from the 34-plex SNP marker set

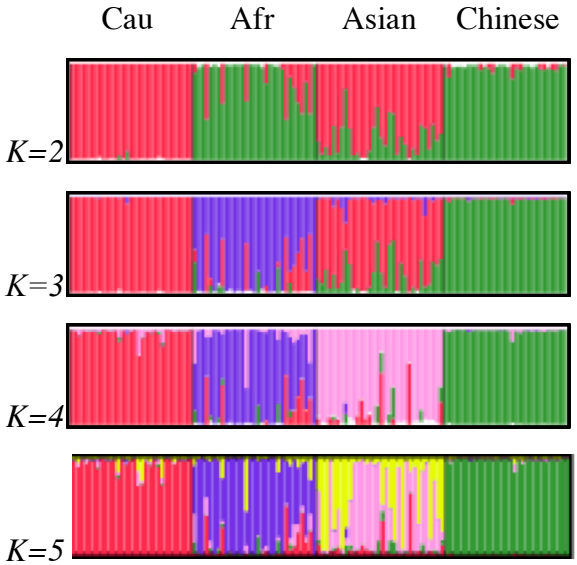


Figure 5.3 *Structure* plot obtained with data from the 36-plex marker set

The data from *Structure* can also be manipulated to produce figures breaking down the affiliation to each inferred genetic cluster at a population level (analogous to that normally produced, but collated per self-declared ethnicity rather than shown individually for each individual). These are displayed in Figure 5.7. From Figure 5.4, Figure 5.5, Figure 5.6, and Figure 5.7 some clear conclusions can be formed.

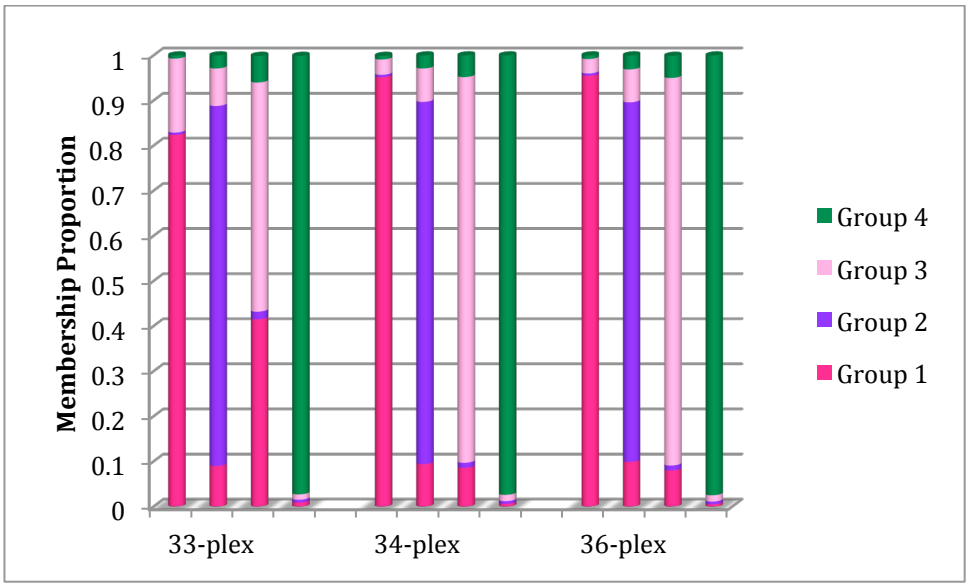


Figure 5.7 Population level structure results for 3 SNP multiplexes
Bars represents the proportions of each theoretically inferred genetic cluster (shown in different colours) within a self-declared population, i.e. the first bar (which happens to represent the British Caucasian population analysed with the 33-plex marker set) shows 82% membership to genetic cluster 1, 16% membership to cluster 3 and under 1% membership to clusters 2 and 4. Each SNP set contains 4 bars corresponding respectively to the British Caucasian, Afro-Caribbean, South Asian and Chinese populations. These results are all for *K=4*, i.e. the model searched for 4 different theoretical genetic clusters within the data.

The genetic proportions ascribed to the Afro-Caribbean and Chinese populations (bars 2 and 4 for each marker set in Figure 5.7) can be seen to remain fairly constant across all three SNP sets (i.e. not surprisingly the addition of extra markers chosen to highlight Caucasian/South Asian differentiation makes little impact on the Afro-Caribbean and Chinese populations). The inclusion of some very powerful Chinese specific markers (e.g. rs3827760) means that the population is separated out very clearly with a 97% membership to group 4 in all three marker sets. The Afro-Caribbean population is not quite as well differentiated (80% membership of group 2 in all marker sets) which could be due to any or a combination of 3 main reasons. It could be that SNP selection was not optimal and that the use of SNPs that can better discriminate between the populations (specifically Afro-Caribbean with Caucasian and South Asian populations) might increase the membership proportions. Alternatively we could be seeing here the genuine results of admixture between populations, and the British Afro-Caribbean population may genuinely have a 9-10% Caucasian and 7-8% South Asian genetic component, at the least in some specific individuals. The structure results in Figures 5.4-5.6 might bear this out as the admixture is largely confined to only a few individuals – this is discussed in more

depth later but may at least explain some of the admixture level. The final explanation may be that we are seeing substructure within the Afro-Caribbean population that is being interpreted here as admixture. Considering the large heterogeneity within Africa this is not unexpected, especially if the 'admixed' samples are coming from North Africa which is known to be genetically distinct from sub-Saharan Africa or from areas like Somalia where geographical factors can influence the genetic makeup of the population, in this case the country's position in the Indian Ocean and proximity to the Middle East and Asia.

A combination of this first (marker selection) and third (population substructure) factors are likely the cause of the admixture seen in the Caucasian and South Asian population in Figures 5.4-5.7. Figure 5.7 clearly demonstrates that increasing the number of SNPs from 33 to 34 (by adding the Caucasian/South Asian discriminating SNP rs16891982) massively increases the differentiation between the Caucasian and South Asian populations, increasing the Caucasian population membership of group 1 from 82% to 95% and increasing the South Asian population membership of group 3 from 51% to 86% (at *K=4*). Hence in this case the addition of an extra SNP showing good discrimination between two populations was sufficient to split the populations better. Adding the extra two Asian SNPs to bring the marker set up to 36 doesn't seem to improve this Caucasian-South Asian differentiation any further at *K=4*, however it is clear from Figure 5.6 that the addition of these two extra markers is able to highlight previously unseen substructure within the South Asian population at the *K=5* level.

At *K=2* (only displayed for the final 36-plex marker combination in Figure 5.6) the Caucasian and South Asian samples cluster together, as do the Afro-Caribbean and Chinese populations. At *K=3* the Afro-Caribbean population is separated from the Chinese population while the Caucasian and South Asian samples still cluster together under all three SNP sets in Figures 5.4-5.6. As discussed above, at *K=4* the South Asian population is well differentiated from the Caucasian population with the 34 and 36-plex marker systems, while this differentiation is much poorer with the 33-plex marker system shown in Figure 5.4. At *K=5* the Structure model is now trying to discern additional genetic structure that is not known to exist with data from only the four populations used. In Figure 5.4 the addition of this fifth group results in

increased admixture levels in the Caucasian and South Asian populations without demonstrating that any genuine genetic signature is being discerned. In Figure 5.5 the fifth cluster is once again failing to find any useful genetic substructure but in Figure 5.6 the addition of the extra two South Asian specific SNPs now allows this fifth genetic cluster to highlight substructure within the South Asian population (in the main, samples within this population now either belong to group 3 or group 5). This elucidated substructure could be the result of religious, geographical, linguistic or cultural (e.g. caste vs tribal) differences. Results for higher values of *K* aren't shown as interpretation becomes progressively more complex with this limited number of samples, however with the 36-plex markers at *K=6* some sub-structuring within the Afro-Caribbean population does become apparent.

5.5 Four Population Classification

The SNP profiles from all four British populations can be used as training sets for the Snipper web-based classification system in an identical way to the results described earlier relating to a three-population classification. Overall cross-validation success rates for the four-population system are 90% with the 33-plex marker set and over 94% with the 34 and 36-plex marker combinations; the increased success with the larger multiplexes due to improvements in the Caucasian and South Asian classification success. The 'apparent success' rate for the 36-plex was over 99%.

Considering the SNP selection criteria were initially focused on finding markers showing skewed allele distributions between two of African, European or East Asian populations, the drop in the classification success rate between a three and four population system, i.e. after adding in the South Asian samples, is not surprising since 33 of the SNPs were not selected with this population in mind. Classification success is still relatively good however – in a similar study by Kosoy *et al* [245] they selected 128 SNPs to define European, African, East Asian and Amerindian populations, and when they added South Asian samples into the mix the South Asian individuals showed 75% membership of an additional population (k=5) with a further 18% membership of the 'European' population. Reduction in the number of SNPs (only using the most informative) decreased this number even more, and if only 48 SNPs were used (still 12 more than in our results presented here) then the South Asian

samples showed 59% membership to the 'South Asian' genetic signature and 29% to the European one. This compares with our figures of 86% membership of the South Asian population and only 8% to the Caucasian one as previously shown in Figure 5.7. More specifically, the Snipper classification results show that this translates to 36-plex cross-validation classification success rates of 90% for the Caucasian individuals, 87% for Afro-Caribbeans and 100% success rates for both Chinese and South Asian individuals (full details are given in Table 5.1). All misclassifying samples were erroneously predicted to be South Asian.

Table 5.1 Population of origin cross-validation classification success using 36 SNP markers

Actual	Predicted			
	Caucasian	Afro-Caribbean	South Asian	Chinese
Caucasian	90%		10%	
Afro-Caribbean		87%	13%	
South Asian			100%	
Chinese				100%

The Bayesian model implemented in the Snipper app suite however is far more complex than just a straight prediction method. Produced for each sample are four likelihood values, each value corresponding to how well that sample fits into one of the 4 defined training set populations. From this data it is possible to see how secure any given prediction is. Figure 5.8 takes the likelihood values produced individually for each sample in the Caucasian and South Asian training sets during the 1-out cross-validation exercise and plots the ratio of the Caucasian/South Asian likelihoods on a logarithmic scale. Any samples above the x axis are predicted to be Caucasian rather than Asian (with increasing certainty) while the converse is true for samples below the axis that are predicted to be Asian rather than Caucasian with increasing certainty. The samples to the left of the y-axis are those self-declared as Caucasian while those to the right of the y-axis are self-declared as Asian. If all samples classify correctly then data points will only be present in the top left and bottom right quadrants of the graph. The likelihood values used are from the 36-plex cross-validation classification exercise, and highlighted in red are the three Caucasian samples that misclassified as

Asian. From this graph it is clear to see that the likelihood ratio between these samples classifying with the Asian training set rather than the Caucasian training set are low, i.e. the three samples may be predicted as South Asian but the model predicts them as only being slightly more likely to be South Asian as Caucasian. By comparison, it is also clear that the LR is far lower than for any genuine Asian sample (the actual LRs for the three misclassifying individuals are 2.4, 2.4 and 6.4 times more likely Asian than Caucasian whereas the lowest similar LR for a genuine Asian sample is 1,193 times with the highest being 96 x 10^{18}). This shows the value of a qualitative system where a prediction is associated with a certainty between two alternate hypotheses.

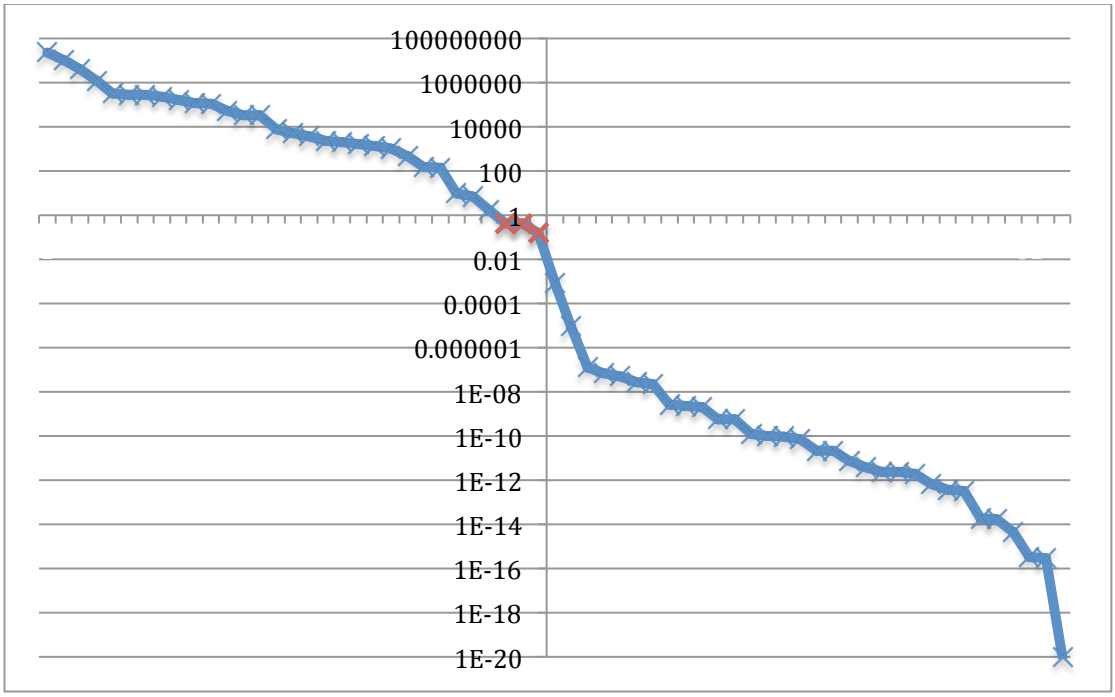


Figure 5.8 Ratio of Caucasian/South Asian likelihood values
Data points to the left of the y-axis represent individuals self-declaring with Caucasian ancestry while those to the right of the y-axis self-declared as South Asian. The Bayesian model implemented in the Snipper prediction app produces likelihood values for any particular sample genetically clustering with samples in pre-defined training sets. The results above arise from the cross-validation classification success test where each sample in a population is removed from the training set and blindly classified, producing likelihood values for membership of each of the four training set populations. A ratio has been taken here between the Caucasian and South Asian likelihood value. A ratio greater than 1 means that the sample is more likely Caucasian than South Asian, while a ratio below 1 indicates the sample is more likely South Asian than Caucasian. The ratios are plotted on a log scale, the further away from the x-axis the stronger the prediction. The ratios from three individuals are highlighted in red, these are samples from Caucasian individuals (as they are to the left of the y-axis), but are predicted to be more likely South Asian than Caucasian by the model (as they are just below the x-axis).

Figure 5.9 present similar data to Figure 5.8, but for the Afro-Caribbean and South Asian samples displaying the ratio between the likelihood of classifying as Afro-Caribbean with that of South Asian. Once again, the four samples misclassifying as South Asian can be seen in the bottom left quadrant, but this time two of the samples have relatively large LRs (although the LR for these samples being Asian rather than Afro-Caribbean is still much less than that achieved for any genuine South Asian sample, so the prediction would still be considered suspect). The LRs of the misclassifying samples are 4.5, 21, 810 and 1,990,706 times predicted more likely South Asian than Afro-Caribbean. On further investigation it was discovered that the 810x LR related to an individual who had actually further described themselves as Somalian, hence the fact that this individual didn't fit very well with the rest of the Afro-Caribbean samples that would predominantly have West African roots. Once again, even though the LR of 810x is quite high, the graph clearly shows that while the sample doesn't cluster with the majority of the Afro-Caribbean samples, it also classifies as more likely Asian with a much lower LR than a genuine South Asian sample (by a factor of a million as the lowest LR for a genuine South Asian sample is still predicted to be over 1 billion times more likely to be South Asian than Afro-Caribbean). This warrants further investigation to determine whether Somalian samples generally would present with this type of likelihood data, and indeed in that case whether an extra Somalian training set could provide realistic differentiation: due to the political instability, a significant proportion of our immigration casework over the last few years has related to Somalian families, suggesting they are a growing demographic within the UK population, and hence this misclassification is an issue needing resolution.

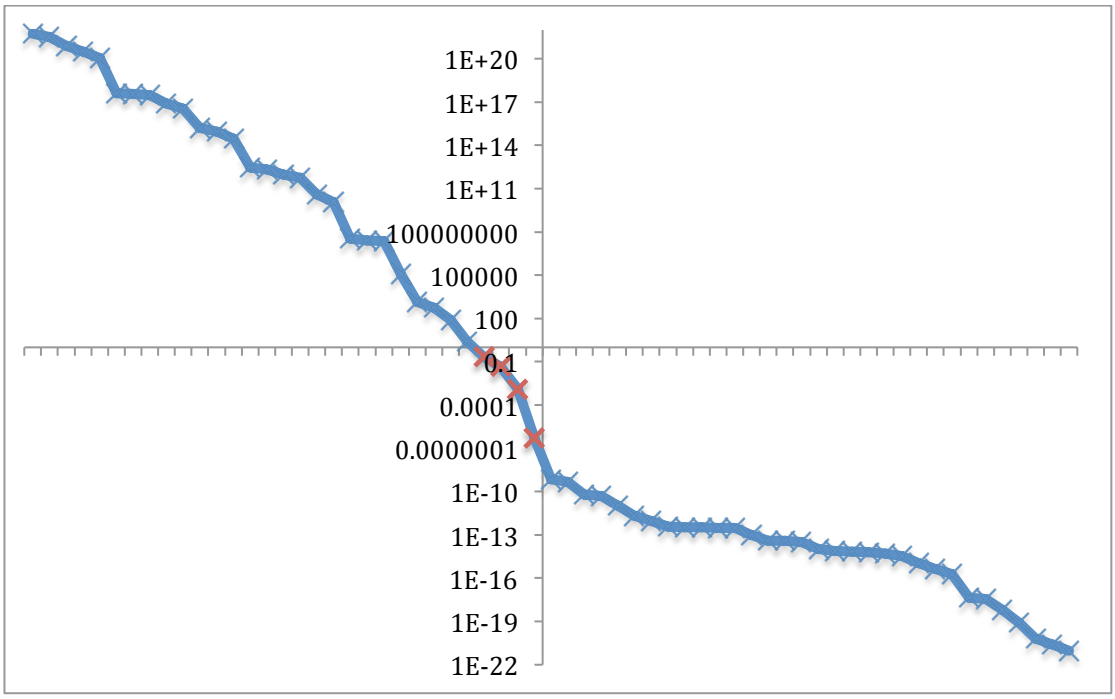


Figure 5.9 Ratio of Afro-Caribbean/South Asian likelihood values
A similar representation of the data to Figure 5.8, but this time taking the ratio of the Afro-Caribbean/South Asian likelihood values. Individuals with self-declared Afro-Caribbean origin are to the left of the y-axis and those with self-declared South Asian origin are to the right. Four Afro-Caribbean samples are highlighted in red that misclassify as South Asian.

The past two figures have presented the outcomes from the 36-plex cross-validation results, however it is also possible to display concurrently the data for the 33 and 34-plex LRs. Figure 5.10 does this for the Afro-Caribbean/South Asian LRs and all 3 marker sets can be seen to produce identical results. This is not the case in Figure 5.11 where the Caucasian/South Asian LRs are displayed (individuals in both populations are ordered by the 36-plex LR, hence this line is smoother): LRs get progressively smaller for the 34 and 33-plex, hence the lines becoming more shallow and closer to the x-axis, demonstrating that while the increase from 34 to 36 SNPs may not affect the overall classification success rate, the addition of two extra Caucasian/South Asian divergent SNPs does help to increase the confidence in the prediction obtained. The overall classification success remains constant between the 34 and 36-plex combinations (while increasing from the 33-plex), however a couple of the specific samples misclassifying can be seen to change. The sample misclassifying with an LR of 6.4x more likely Asian than Caucasian with the 36-plex is correctly predicted to be Caucasian with the 34 (and indeed the 33) marker sets in Figure 5.11. The alteration to a very weak South Asian classification with the extra

SNPs is due to the GG genotype at rs2715883 that is not otherwise observed in any Caucasian individual tested, hence adding uncertainty to the prediction.

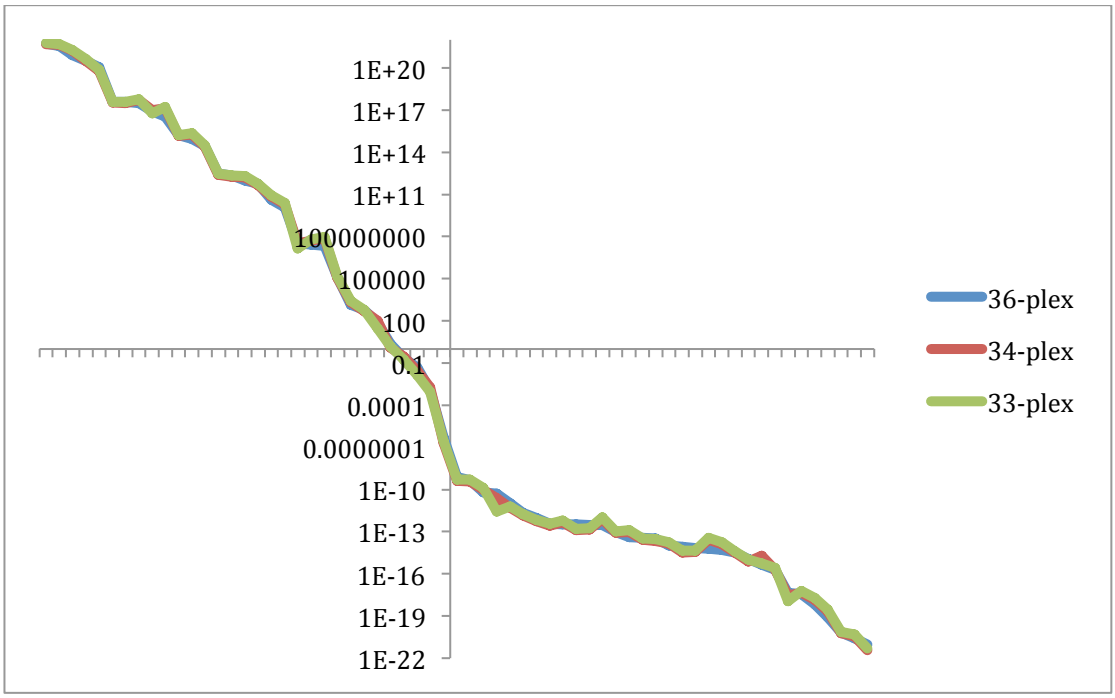


Figure 5.10 Ratio of Afro-Caribbean/South Asian likelihood values obtained with three different marker sets

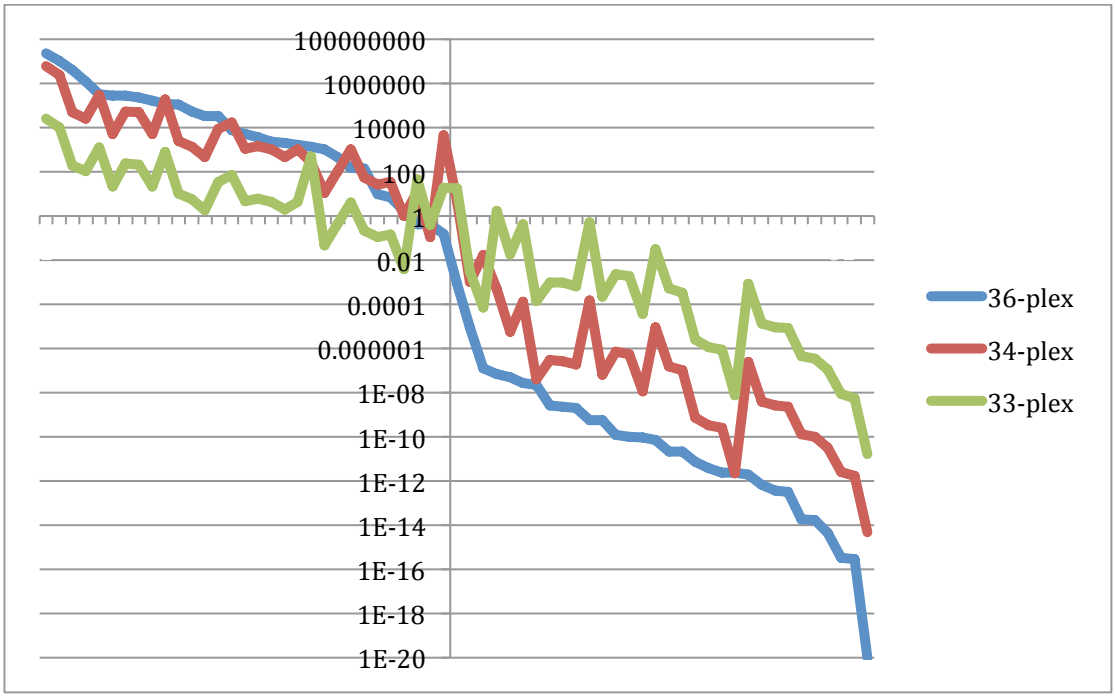


Figure 5.11 Ratio of the Caucasian/South Asian likelihood values obtained with three different marker sets

Finally, a subset of 20 of the incorrectly classifying samples from either the mitochondrial haplogroup or Y-STR prediction systems were typed for the 36-plex population specific SNPs. All samples bar one classified correctly with the SNPs, and the strength of these classifications is shown in figures 5.12-5.16. The training sets previously produced for the four populations (consisting of a total of 125 individuals) were used to calibrate the Bayesian model, and then the new samples to be tested were run through this prediction system. Two-way population classification likelihood ratios are displayed in the figures showing the cross-validation results for each sample from the relevant training sets as well as where the actual results from any applicable tested samples sit within the range obtained from the training sets.

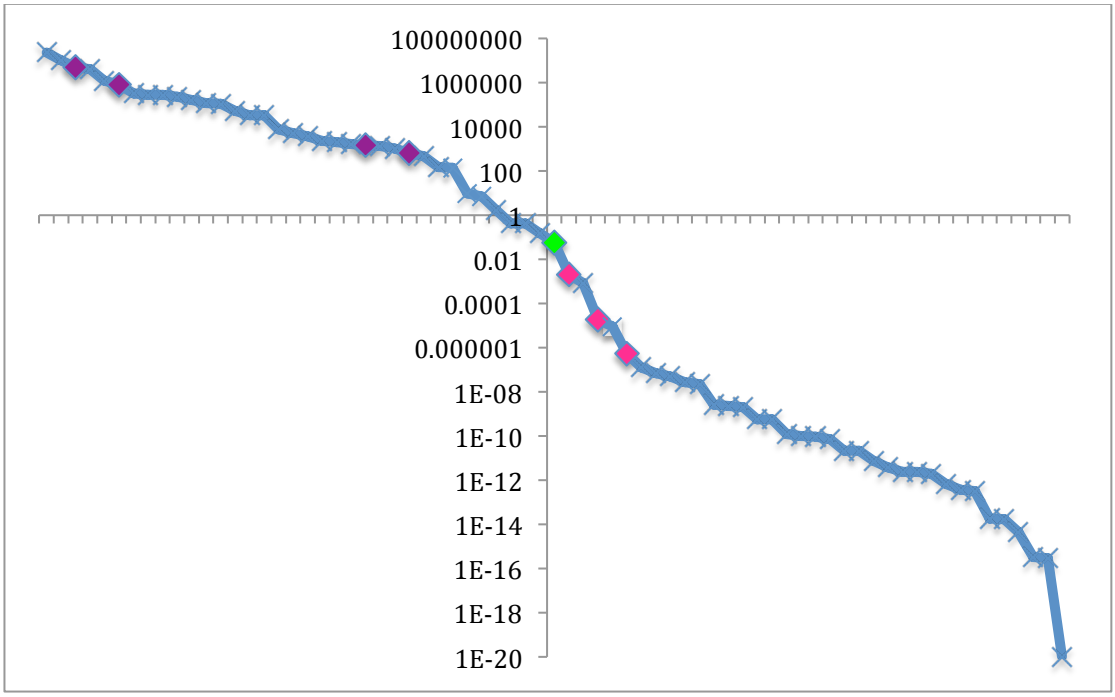


Figure 5.12 Ratio of Caucasian/South Asian likelihood values for a selection of previously misclassifying samples
Likelihood values obtained from the one-out cross-validation success estimate were used for the Caucasian and South Asian training sets. Caucasian training set samples are located left of the y-axis represented with a blue cross, while South Asian training set samples are similarly located right of the y-axis. Samples predicted to be more likely Caucasian than Asian are positioned above the x-axis with those more likely South Asian than Caucasian below the x-axis. Likelihood ratios are displayed on a logarithmic scale – the further away from the x-axis the stronger the prediction. The Bayesian model implemented in Snipper and trained with the previously described four population training sets was used to classify eight samples. All eight samples had previously been misclassified by either the Y-STR or mitochondrial haplogroup prediction system. Individuals highlighted in purple self-declared as Caucasian but had previously been predicted to be Asian (and here are all correctly predicted to be Caucasian using the 36-plex SNPs). Individuals highlighted in pink self-declared as South Asian but previous predictions with alternate systems had classified them as Caucasian (and here are all predicted to be South Asian with the SNPs). The sample highlighted in green is correctly predicted here to be South Asian while had previously been predicted to be of Afro-Caribbean origin with the Y-STRs – the SNP results give a ratio of only 18 times more likely that this sample is South Asian than Caucasian.

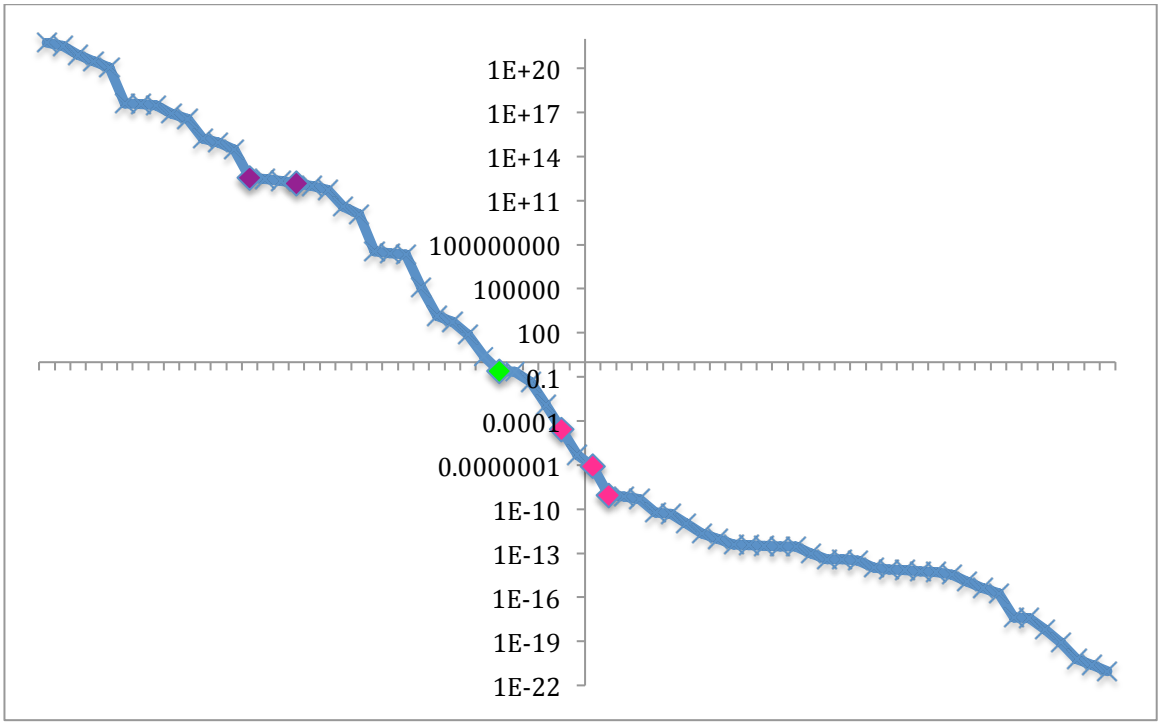


Figure 5.13 Ratio of Afro-Caribbean/South Asian likelihood membership values with a selection of previously misclassifying samples
Ratios for each sample in the UK Afro-Caribbean training set are displayed as blue crosses to the left of the y-axis while those for the UK South Asian samples are displayed as a blue cross to the right of the y-axis. Predictions more likely Afro-Caribbean than South Asian are displayed on a logarithmic scale above the x-axis while the reverse is true of points below the x-axis. Samples in purple are Afro-Caribbean individuals that had previously been misclassified as South Asian using mitochondrial DNA or Y-STRs while the opposite is true of those individuals in pink who describe themselves as Asian but had previously been erroneously predicted to be Afro-Caribbean. All five of these samples are correctly predicted here using the SNP system, the purple ones very strongly while the pink ones relatively weakly in comparison to the training set samples. The Afro-Caribbean sample in green had previously been predicted as Caucasian using mitochondrial DNA and is here also incorrectly classified, though not as either Caucasian or Afro-Caribbean but instead as 4 times more likely to be South Asian than Afro-Caribbean.

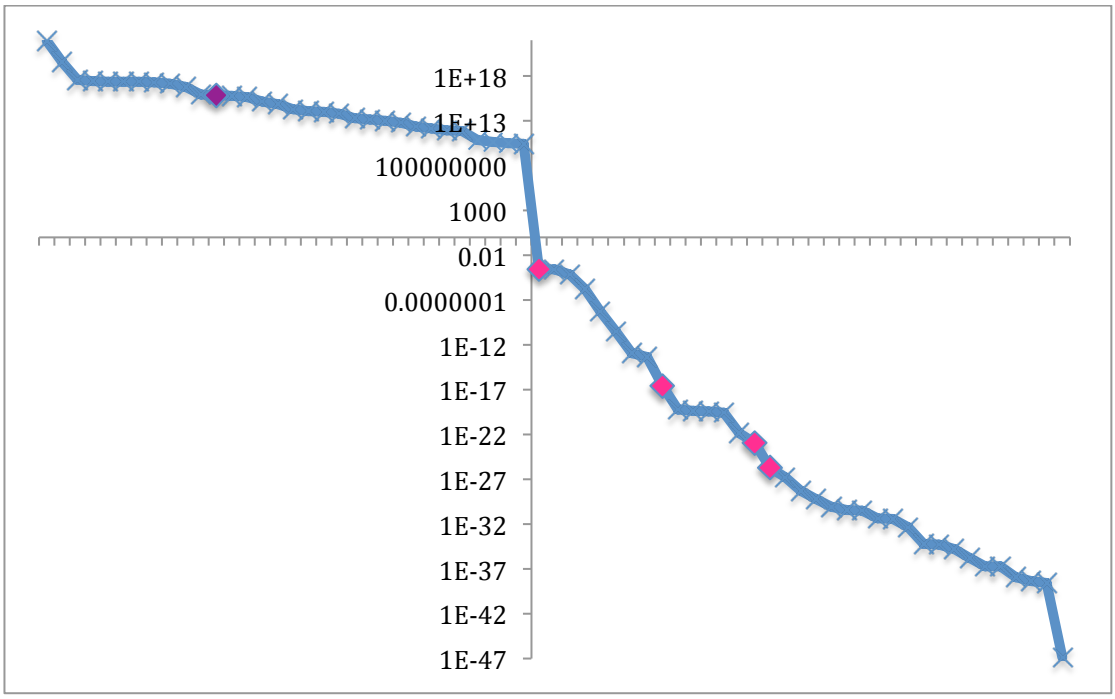


Figure 5.14 Ratio of Caucasian/Afro-Caribbean likelihood values for a selection of previously misclassifying samples
The Caucasian training set samples are to the left of the x-axis and ratios denoting a greater likelihood of a sample belonging to the Caucasian population rather than the Afro-Caribbean population are plotted above the x-axis. To the right of the y-axis with blue crosses are the Afro-Caribbean training set samples and points below the x-axis correlate with LRs more likely to be Afro-Caribbean compared to Caucasian. The Caucasian individual highlighted in purple had previously been predicted to be Afro-Caribbean with the mitochondrial DNA but here is correctly classified as Caucasian while the opposite is true of the four pink Afro-Caribbean samples that had previously been wrongly predicted as Caucasian but are here correctly predicted as Afro-Caribbean.

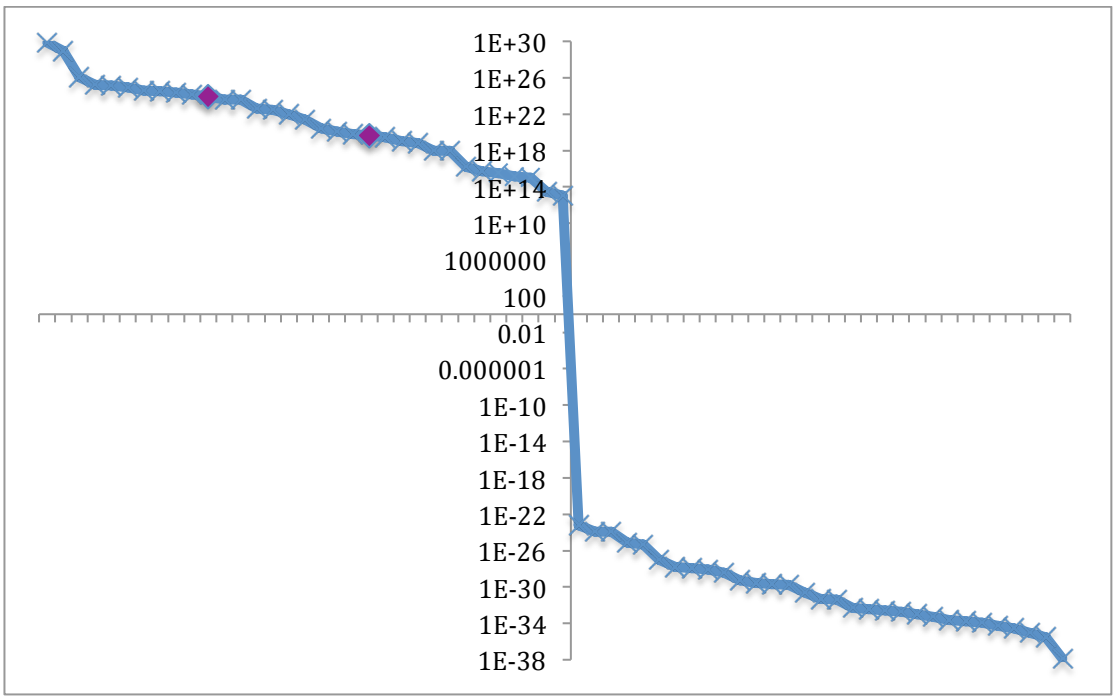


Figure 5.15 Ratio of Caucasian/Chinese likelihood values for a selection of previously misclassifying samples
As previous figures but here the Caucasian training set is to the left of the y-axis and the Chinese training set to the right. Purple samples represent Caucasian individuals previously misclassified as East Asian based on mitochondrial haplogroups – here correctly classifying very strongly as more likely Caucasian than Chinese.

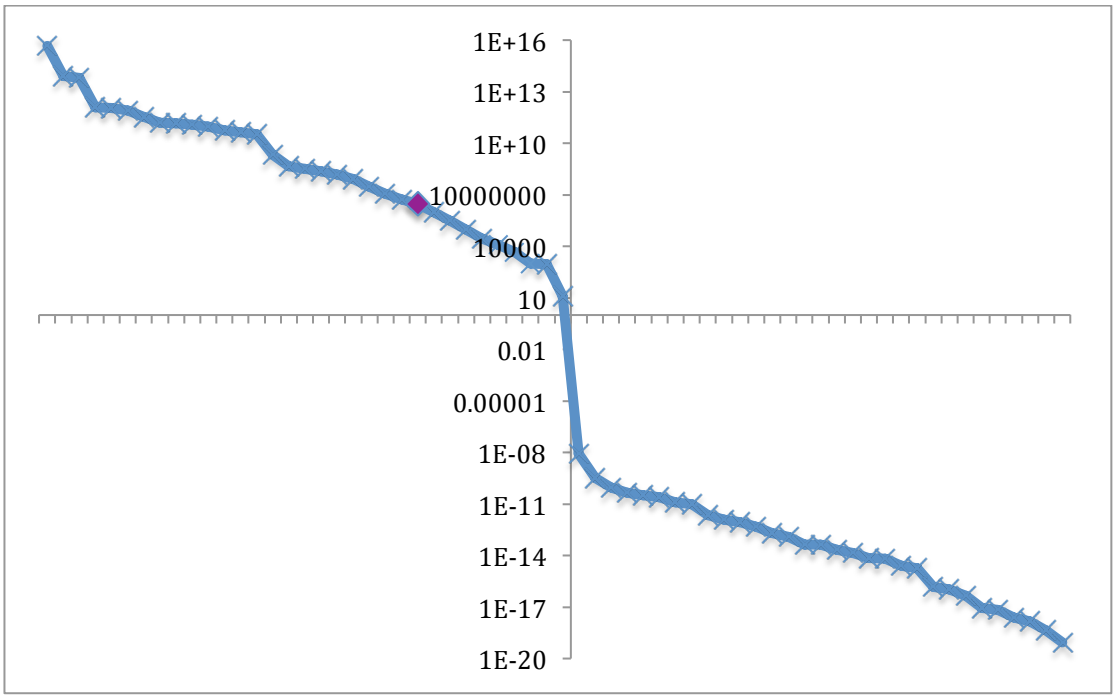


Figure 5.16 Ratio of South Asian/Chinese likelihood values
Similar to the previous four figures, except the South Asian training set samples are here displayed to the left of the y-axis and the Chinese training set samples to the right. The South Asian sample in purple had previously been predicted to be East Asian from the mitochondrial results but is here correctly classified.

6 Final Discussion

6.1 Populations

The analysis of this diverse set of genetic markers within the British population has drawn out some consistent themes regarding the genetic composition of the different sub-populations. The British Caucasian population is shown to have less diversity than either the Afro-Caribbean or South Asian populations with both the mitochondrial and Y-STR data. Human evolutionary/migratory events are known to have resulted in elevated genetic diversity levels within Africa and Asia due to the age of, and subsequent variation within, the founding lineages, hence Western Europeans are known to generally have less diversity. Having said that, the violent history of Britain littered with numerous conquest and invasion events might also suggest that this population should be more diverse than some other more stable ancient populations. This is shown to be true in comparison to the Irish population, which shows both less diversity, and at the same time elevated frequencies of specific rare worldwide types, for both the Y-STR and mitochondrial data. For some British Caucasian samples, the haplotype or sequence associations with other areas of the world seem to provide a hint of the genetic legacy brought about by the complex history of the British Isles.

The high diversity within the Afro-Caribbean and South Asian populations is also a result of the wide geographical areas from which these individuals ancestrally trace back to, yet all this diversity is grouped together as one overarching sub-population within Britain. The sub-structuring in the South Asian population is highlighted by the 36-plex population specific SNP structure results at K=5 (Figure 5.6) where a Bayesian model is able to detect multiple genetic signatures within 'one' population. The variation in cultural practices, religion, language and geographical boundaries in the Asian sub-continent would suggest that the UK South Asian population might indeed be expected to show significant stratification. This stratification is additionally reflected in the Y chromosome (Figure 3.16), and especially mitochondrial (Figure 4.16), results presented here where large intra-population variation is observed with some samples clustering close together, and some further

away. Interestingly, an analysis of the British Asian population with standard autosomal forensic STRs shows that the distribution in UK Hindu/Sikh individuals is very similar to that in UK Muslim individuals [246].

Analysis of SNP data in India from different subpopulations has raised the possibility of using this substructure to further refine the South Asian classification [235]. Using just 12 SNPs the Indian Genome Variation Consortium was able to achieve a 100% classification result between tribal and non-tribal populations, a classification into the four main linguistic groups could be achieved with over an 85% success rate using 8 SNPs, and a 6 region Indian geographic classification could be achieved with a 55% success rate using 6 SNPs [235]. It could be postulated that use of these SNPs may not just allow sub-categorisation of the Indian population, but may also provide some information to facilitate a partial classification between different South Asian populations (due to differing linguistic, tribal and geographic distributions), however time constraints meant that this could not be tested.

For the Afro-Caribbean population, mitochondrial haplogroup analysis provided a very good differentiation between African and non-African ancestry as principally showed by membership to haplogroup L: 96% of UK Afro-Caribbean samples along with 96% of Caribbean island samples presented with an L haplogroup while of the remaining 302 Irish, British Caucasian and South Asian samples only 1 possessed a haplogroup L sequence. The logical interpretation of this disparity is that the British Afro-Caribbean population contains a maximum possible non-African maternal admixture rate of only 4%, while on the Y-STR system a full 33% of Afro-Caribbean samples didn't classify correctly. This may reflect a much larger non-African male admixture component in the British Afro-Caribbean population in comparison to the female level, or else just suggest that the Y-STR classification system couldn't cope with the large variation in genetics across Africa – quite possibly it is a mixture of the two. The SNP analysis of the Afro-Caribbean population should be able to quantify overall admixture levels more accurately, and as shown in Figure 5.7 the Afro-Caribbean population is found by this method to have contributing genetic fractions of 10% Western European, 7% South Asian and 3% East Asian.

The level of admixture in the UK Afro-Caribbean population is expected to be shaped by the population history, including the proportions of individuals tracing back their ancestry directly to Africa or *via* the Caribbean, where the slave trade had a major impact. The slave trade also played the crucial role in the African-American population development, and comparison with admixture results in this population is instructive.

During the 18th, and the beginning of the 19th century, the Atlantic slave trade is estimated to have resulted in the transportation of 380,000-570,000 slaves from Africa to America [247, 248]. This genetic inflow provided the foundation for the present African American population group. Numerous different research projects have studied the genetic ancestry of present day African-American populations [142, 210, 215, 216]. A large scale autosomal STR and in-del typing study using *Structure* analysis [142] ascertained that African American populations from North Carolina, Baltimore, Chicago and Pittsburgh showed a mean of 69-74% ancestry from an African ancestry cluster predominantly associated with West Africa and 11-15% European/Middle Eastern ancestry, which is a similar European proportion to that revealed in our British Afro-Caribbean population with the SNP work. The proportion of African and European ancestry varied considerably between individuals, a reflection of the non-homogeneous nature of individuals self certifying as African-American. The mean European contribution calculated using these markers is broadly in line with other studies, although some outlier populations are known to exist (e.g. the African-American 'Sea Islanders' of South Carolina and Georgia who are believed to have a much lower European influence of only 3.5%) [215]. Further analysis by Tishkoff [142] was able to posit that for most African Americans the African ancestry portion is likely to be mixed, containing contributions from different areas of Western Africa. We are currently carrying out further work analysing some distinct Western and Eastern African populations to see if we can make any similar inferences about the Afro-Caribbean population, however given the history of the slave trade and the known demographic makeup of migrants arriving in the UK directly from Africa, it is expected that a similar overall West African pattern would be discerned.

6.2 Classification Systems

In both the Y chromosome and mitochondrial DNA classification system, very good success rates were obtained for predictions of Afro-Caribbean origin, while the more significant problem has been trying to fully separate the Caucasian and South Asian samples: in the Y-STRs the final step in the classification system left 77 samples of which 41 were South Asian and 22 were Caucasian, while for the mitochondrial DNA there was good classification success rate (96%) for the Caucasian samples but only 77% correct classification for the South Asian individuals with 17% erroneously being classed as Caucasian. Separating the South Asian and Caucasian individuals with the population specific SNPs also entailed much more challenging marker discovery and selection than finding SNPs with good Caucasian/African/East Asian divergence. The gene flow east into Asia from western Europe is undoubtedly a major factor in this classification difficulty, and the prevalence of Western Eurasian mitochondrial haplogroups is known to be greatest in Pakistan and Western India rather than further south and east [231], hence the error rate may be higher for a UK South Asian individual of Pakistani heritage rather than Bangladeshi. This migration was once thought to have taken the form of an large invasion about 4,000 years ago but is now known to have happened either gradually or in multiple small scale events over the last 10,000 years [249], possibly corresponding to the domestication of various crops and expansion of agriculture [228, 249]. Given the opportunity the subsequent passage of time provides for sequence divergence, it is undoubtedly true that a more detailed mitochondrial classification system focussing more on sub-lineages (which would have evolved more recently) would improve the classification figures. Similarly for the Y chromosome work, while it became very hard to separate out the Caucasian and South Asian samples towards the end of the decision tree using the more rapidly evolving STRs, it is quite possible that had slowly evolving Y-SNPs been used a better result might have been obtained here since it is to be assumed that genuine recent admixture isn't going to be the cause of all this error, but merely an inability to separate the populations using those haplotype markers.

Overall the 34 or 36-plex population specific SNP marker set gave the best results, with a predictive success rate of over 94% when judged using the stringent one-out

cross-validation assessment method (or 99.2% if using 'apparent success'), and additionally this system differentiated four populations (including Chinese) unlike the uniparental markers (Y & mitochondria) which were only used to distinguish between three British populations. Classification success rates with the mitochondrial haplogroup method was 90% and with the Y-STR system a slightly less impressive 80%. In all of these methods, either a more detailed classification system (mitochondria) or use of a 'not-determined' group for prediction calls with low certainty (Y-STRs and population specific SNPs) would have improved these success rates. One drawback with using uniparental markers is that this will not discern admixture in individuals of mixed race since all the genetic information is coming from a single parent. While no SNP data for mixed race individuals has been included here, preliminary work has been carried out and when included with the four training sets and visualised with *Structure* the mixed ancestry is generally very clear (i.e. these individuals show 50:50 membership to two different groups). It is assumed that if *Structure* can pick up these two genetic signatures then the Snipper app will also be able to and individuals will classify with fair strength into two genetic clusters and when plotted on a graph such as Figure 5.14 the sample will place near the x-axis between the two training sets, but this has yet to be fully tested.

Interestingly each of the three systems has a tendency to classify different populations best, partly due to the different way that these systems deal with hard-to-classify samples (i.e. which population they seem to go to by default). The Y-STR system classifies the South Asian samples best, with a correct South Asian prediction in 90% of cases, while the mitochondrial system is the opposite with good success rates for Afro-Caribbean and Caucasian samples – individuals from both of these populations successfully classify 96% of the time. With the populations specific SNPs, the South Asian and Chinese samples classify best, with 100% correctly classifying, while similar to the Y-STRs most samples that misclassify are wrongly predicted to be South Asian. This shows the value of using more than one prediction method.

A selection of misclassifying samples were re-typed with other systems with the aim of seeing whether the misclassification could be explained. Mitochondrial sequencing was performed on 42 samples that misclassified either with the Y-STR or SNP system while Y-STR results were available for 18 samples that misclassified under the

mitochondrial DNA prediction system (only male samples could be re-tested for the Y-STR system). As reported in section 5.5 population of origin prediction was carried out using the 36-plex SNP marker set on a further 20 individuals that had either misclassified with the Y-STR or mitochondrial DNA results.

Of those 18 samples that were tested with Y-STR markers having misclassified using mitochondrial DNA analysis (all were South Asian samples that had been predicted to be Caucasian or 'unknown'), all correctly classified under the Y prediction system. This may suggest fairly recent mixed ancestry in these individuals (a Caucasian maternal influence on the background of a largely South Asian genetic component) but there is nothing historically relevant to account for such a large number of admixed individuals (as opposed, for example, to the situation in America with the Atlantic salve trade). More likely it is just a factor of the known eastern movement of typically European genomes into the western areas of Asia in the last 10,000 years as discussed above. A more refined mitochondrial haplogroup prediction system including finer sub-lineage differentiation is unlikely to correct all of these misclassifications, but would undoubtedly improve the success rate since the South Asian ancestry can be correctly predicted from the male lineage proving there is sufficient genetic divergence at some level: if alternatively the misclassifications for both mitochondrial and Y chromosome analysis had been coincidental in the same samples, then it would have had to be assumed that there was a particular subpopulation within South Asia that would be very hard to distinguish from Caucasians.

Of the misclassifying Y-STR samples analysed with mitochondrial DNA, all 35 gave a correct ancestry prediction from the updated mitochondrial haplogroup system. One of the Afro-Caribbean samples predicted to be South Asian from the Y chromosome results possessed a M1a1 mitochondrial haplogroup which does correctly predict to be African, but suggests a Northeast African ancestry rather than sub-Saharan, hence possibly why the Y-STR prediction was wrong. Two South Asian samples that were predicted to be Caucasian from the Y results both possessed U7 mitochondrial DNA haplogroups. The updated classification system now includes the U7 lineage in the South Asian predictors – these two samples demonstrating how useful the more refined classification criteria are. As previously discussed, U7 is seen at high

frequencies in Iran and Pakistan, and hence these two individuals are most likely of Pakistani descent, and therefore amongst the hardest of the South Asian samples to separate from Western Europeans.

The advantage of re-typing misclassifying Y chromosome or mitochondrial samples with the population specific SNPs is that in addition to a prediction, the result obtained also gives a qualitative value on the prediction strength. Figures 5.12-5.16 plot the results of these SNP predictions and put the strength of the prediction into context with respect to the samples of known origin used in the training sets to 'teach' the Bayesian classification model. Three samples from the mitochondrial study with unusual results that predicted them to be of East Asian ancestry can be seen to classify well with the 36-plex SNP test; not only are they correctly predicted, but the confidence in the prediction is also strong (see Figure 5.15 and Figure 5.16). Figure 5.14 shows that the single Caucasian sample to present with an L (African) mitochondrial sequence classifies very strongly as Caucasian with the SNPs, while of the four Afro-Caribbean samples re-tested (two had previously been misclassified as Caucasian with Y-STRs and two with mitochondria), all four correctly classify as Afro-Caribbean rather than Caucasian. One of these samples can be observed to only classify very weakly as more likely Afro-Caribbean than Caucasian having been predicted as Caucasian from the fairly atypical Caucasian haplogroup HV1b1, and indeed it is a little misleading to suggest this classifies correctly as Afro-Caribbean rather than Caucasian because even though that is true, this sample is also plotted in green in Figure 5.13 because it actually fits best with the South Asian training set, and is found to be 4 times more likely South Asian than Afro-Caribbean. This is the one sample out of the 20 re-analysed with the SNPs that still misclassifies, although the LR here is so weak that the prediction is essentially 'undetermined', and indeed while this sample is wrongly classified as South Asian in Figure 5.13, from the LR it is still shown to be a 'more' Afro-Caribbean than 3 of the 31 samples in the Afro-Caribbean training set. Further investigation revealed that this sample belonged to an individual who had additionally classified themselves as Somalian rather than just Afro-Caribbean.

In Figure 5.12, of the three Asian samples in pink that had previously been wrongly classified as Caucasian with the mitochondrial results all correctly predict as South

Asian with the SNPs but in each case with relatively low likelihood ratios of being Asian rather than Caucasian. The combination of Caucasian mitochondrial haplogroups and weak (correct) South Asian SNP prediction most likely indicates that these samples are from the western periphery of the Asian sub-continent and not separating as clearly from the Caucasian samples. More detailed knowledge of different subpopulations within South Asia, for the SNPs, mitochondria and Y-STRs, may yet allow a slightly more detailed classification to be obtained if these consistently weak predictions do relate to a west-east genetic cline (i.e. a weak South Asian classification might indicate Pakistani origins while a strong classification may indicate Bengali or Bangladeshi ancestry).

The use of the population specific SNPs here is shown to be of high evidentiary value since not only did they correct 19 of the 20 Y-STR or mitochondrial misclassifications, but by providing a measure of confidence in the classification it is also possible to draw more firm conclusions, i.e. of the 20 samples at least 14 are shown to classify with a high degree of certainty, while as many as 6 (including the sample that misclassifies) would be considered problematic predictions on the basis of this data. A simple rule stating that any prediction will be considered unsafe if the classification is weaker than the weakest sample in the training set would immediately have removed all of the misclassification events, i.e. four likelihood values are given for the likelihood that the tested sample fits into each of the four training set populations, if the ratio of the two best likelihood values is taken (i.e. the values for the two populations where the sample is most likely to belong) then if this ratio is worse than all the ratios achieved from the training set for 'known' samples then the classification would be considered unsafe – this effectively adds an 'unknown' zone to the centre of the graphs in Figures 5.8-5.16.

Of the seven samples shown in section 5.5 to have misclassified with the SNPs on the basis of the cross-validation assessment of prediction success, the three Caucasian samples that were very weakly predicted to be South Asian (Figure 5.8) all correctly classified as Caucasian with mitochondrial DNA. The four misclassifying Afro-Caribbean samples are more intriguing. The sample predicted to be 800x more likely South Asian than Afro-Caribbean, and subsequently discovered to be Somalian, possessed an L2a mitochondrial sequence predicting maternal African ancestry. The

two samples predicted relatively weakly to be South Asian rather than Afro-Caribbean (4.5x and 21.3x) had respectively T1a and H* mitochondrial haplogroups which would both cause these samples to be classified as Caucasian from the mitochondrial sequence. That means that the two prediction methods have given different answers, neither of which was correct. The weakness of the SNP classification means that it wouldn't be trusted (the lowest South Asian/Afro-Caribbean likelihood ratio for a genuine South Asian training set sample was over 1 billion as shown in Figure 5.9 which is very different to 4.5 or 21.3), however obviously these two Afro-Caribbean samples are very atypical of the Afro-Caribbean data set as a whole. The mitochondrial analysis of the individuals from the Caribbean showed that all 44 tested Jamaican samples belonged to haplogroup L along with 52 out of 56 from other assorted islands. More work would have to be carried out on assessing different distinct populations within Africa (which we are currently doing) however it appears likely that these two Afro-Caribbean samples classifying maternally as Caucasian and weakly with the SNP set as South Asian, most likely trace their ancestry back to northern or eastern Africa in communities with a higher incidence of non-sub-Saharan gene flow.

The only misclassifying sample from the 36-plex SNP analysis that might conceivably be reported as belonging to the incorrect population would be N25 which is predicted to be South Asian with an LR of nearly 2 million for being South Asian rather than Afro-Caribbean (although even here, this LR is still nearly 700 times less than the lowest similar LR observed in the South Asian training set, so if adding the 'unknown' zone to the graph then this prediction would be considered insecure). Sample N25 is from 1999, classified as Black at the time, and there is no way to verify this data now since the sample is annonymised. The Y-STR type comes out as Afro-Caribbean when put through the classifier system due to the presence of a 21 allele at DYS390, however many of the other markers possess alleles seen very infrequently in this population, e.g. generally DYS385 alleles are quite long in UK Afro-Caribbean individuals as shown in Figure 3.13 but this sample has a genotype of 11-12. Searching this haplotype on the YHRD produces no matches to any of the 76,613 worldwide samples typed for these Y-STR markers, which is fairly unusual. The conventional protein system Gc was used in the past for identification purposes before the advent of DNA testing, and the phenotypes here can show different

geographical distributions with the 1F phenotype being far more common in Africa than Europe. Unfortunately this sample types as a 1F1S phenotype which is found at roughly the same frequency in both a UK Caucasian and UK Afro-Caribbean population (17% vs 20% - laboratory frequencies from our unit when we used to do this test for relationship cases). The mitochondrial haplogroup is T1a, which is a European type and hence corroborates the data from the SNP study that the sample is not Afro-Caribbean, but now suggests it may be Caucasian rather than Asian. Re-extraction of the sample produced identical results and showed that sample mix-up was not a causative factor in the misclassification. The conclusion from all this data is that sample N25 is certainly not a typical sub-Saharan African or UK Afro-Caribbean sample and the results are probably down to one of four possibilities:

a) The individual comes from a remote sub-Saharan African tribe that has been genetically isolated and hence has different genetic characteristics to other African samples. This seems unlikely given the fact that it differs significantly at more than one marker – the mitochondrial DNA does not belong to the L sub-type from which all sub-Saharan African populations are known to derive so this isolated genetic group would have to be founded by a European woman, and the SNPs also show non-African influences.
b) The individual comes from a different population group completely, such as North Africa, hence the Ys classify as African but the mitochondria and SNPs show a non sub-Saharan African origin. The individual in this case could still have been classified as Black British at the time of sampling.
c) The individual may be of mixed race with a Caucasian mother and Afro-Caribbean father. Further research on the SNP classification (data not included here) has shown this to be unlikely since over 90% of mixed individuals classify weakly as Afro-Caribbean due to the fixed nature of some African selective SNPs where the African allele is never normally seen in European populations.
d) The final possibility is that it could genuinely be a UK Caucasian or South Asian sample that has been mis-labelled at the sampling stage, although considering that the three different classification methods have all predicted different ancestral origins there is no particular reason to believe that this is a likely option.

The most likely of these four scenarios would seem to be (b), and in all likelihood the confusing classification of this sample just means it doesn't fit well into any group

and is most likely of Northern/Eastern African origin or indeed even from the Near East or Middle East (Figure 1.14 highlights the genetic diversity between areas of Africa). By relaxing the search parameters on the Y haplotype reference database, two somewhat similar haplotype can be found in the database: an Arab individual from Morocco where the DYS437 allele is different by 2 repeats, and an individual from Lviv in Ukrane where the DYS391 allele differs by 1 repeat. This would appear to provide further supporting evidence that theory (b) is the most likely.

6.3 Impact of work

There has been one previous attempt at population-of-origin classification of an unknown DNA sample within a UK population. In 2001, research by the Forensic Science Service in the UK tried to separate out the main UK ethnic populations on the basis of 6 routinely used STRs in criminal testing [250]. The rational for this was the known variation in allele frequency between populations for these STRs. The ethnicity of the samples in the study was determined visually by a police office rather than from any known ancestry information, hence is slightly different to those samples used throughout this project where ethnicity was self-reported. Additionally they use a 5-population classification system including the Middle East; this is likely to negatively influence the predictive success due to the geographical proximity of Africa, Europe and Asia to this region. Success rates were 56% for Caucasians, 67% for Afro-Caribbeans, 43% for South Asian, 66% for Southeast Asians and 30% for individuals classified as Middle Eastern. The same Caucasian/South Asian misclassification effect was seen with 15% of Caucasians predicted as South Asian and 17% of South Asian individuals predicted to be Caucasian (with an additional 18% predicted to be Southeast Asian). Overall this classification system manages to achieve only a 52% success rate, and while the inclusion of the Middle East will undoubtedly have contributed to this reduced value, it is clear even at the individual population level that there is a large error rate generated when using this STR based solution. This compares with conservative success rates in this study of 80% for the Y-STR method, 90% for the mitochondrial DNA analysis and over 94% for the SNP classification system, highlighting the increased utility of any of these methods over the previously suggested STR method.

The application of this research to live casework also has other associated issues that need to be addressed in a wider ethical, legal and social context. In a special Nature Genetics issue on 'race' in 2004, Cho and Sankar, two biomedical ethics specialists, outlined their concerns regarding the implications of population classification in forensic genetics [251]. The concept of race first came about when European explorers journeyed across the globe, but research in medical genetics has questioned the utility of applying diffuse and changeable social labels such as race to genetic variation (at least with regard to disease susceptibility and drug efficacy) [252]. While noting that classifying people according to race/continental origin can have very little benefit inasmuch as broadly delineating genetic variation across the world, and hence questionable benefit for medical genetics, there are still specific genetic markers that do show significant population differences as acknowledging by the authors [252] (who use skin colour as their example), even if Cho and Sankar [251] choose to infer a different conclusion when referencing that paper. The argument put forward by Cho and Sankar is that there is so much variation with how people self-describe themselves, that the diversity within an ethical or racial group can encompass such a wide spectrum of individuals, and that the boundaries between different ethnicities can be very blurred, that it is not useful to label people as such [251]. Furthermore, they note that physical appearance and ancestry can be quite different. The diversity within a group can be seen clearly with the mitochondrial results presented in chapter 4, and indeed physically if comparing opposing ends of a continent (e.g. Finland and Spain, Sri Lanka and Pakistan or Nigeria and Somalia), in agreement with their view that ill-defined labels can be responsible for artificially grouping together quite disparate people. The differences between the continental (or sub-continental) groups however, would still appear to be sufficient for a rough classification to be viable, as proved by the success rate with the SNP classification system that shows it is possible in the UK to distinguish between the four major ethnic groups successfully on the basis of self-assigned ancestry - although obviously this doesn't take into account the small proportion of individuals that fall outside these four groups or come from areas on the boundaries between them.

The acknowledgment that genetics can be used to broadly separate individuals into different populations/ethnicities doesn't imply that there are major genetic differences

between them and that the concept of 'race' has a large genetic foundation (there is much more variation within populations than between, and as already stated, medical studies have shown the fallacy in using simple continental origin as a proxy for detecting genetic differences) but rather that as populations have moved and changed, specific genetic signatures can be detected, in a similar way to how regional accents differ across a country.

Further issues are raised in a letter to the editor of Forensic Science International: Genetics about whether progressing with these types of tests would lead to discrimination and stigmatization of a specific community once results show that a perpetrator is likely to belong to that grouping [253]. This is obviously a serious issue, and something that must be considered before the release of such data into the public realm (e.g. as part of a police appeal for information), however in current form it is primarily anticipated that this technique will be used to provide intelligence to the police about which suspects to prioritise rather than be part of a 'genetic photo-fit' since there is still some degree of uncertainty in any prediction (especially if an individual comes from a population not included in the prediction model, e.g. Pacific Islanders). In addition, the information gained from this 'molecular eye-witness' is no different to that from a normal eye-witness, hence exactly the same arguments would apply to ever using normal eye-witness testimony, and yet eye-witness descriptions and photo-fits are routinely used in police work to aid the apprehension of individuals implicated in criminal behaviour, and this does not routinely lead to public disorder or mass genetic screening of specific communities as suggested, although that doesn't mean we should dismiss the potential for this to occur in specific cases. Furthermore the molecular eye-witness might be more accurate than real eye-witness testimony which is known to have serious flaws [254]: over 300 people in the US convicted of a serious crime, including 18 on death row, have been exonerated by DNA re-examination of the evidence, and erroneous eye-witness testimony was a contributing factor in 72% of those cases [255]. The SNP classification system can provide a known predictive success rate and any suspect in the crime would still have to go through the normal DNA profiling procedures excluding an innocent individual. In reply to this letter, Kayser *et al* [256] also make the point that eye-witness statements are essentially no different while additionally noting that there are also other issues associated with ethnic prediction to be debated,

specifically that if the testing unmasks an unknown ancestry component, e.g. that the individual is likely to be of mixed race, does this violate the individuals right-not-to-know (i.e. they now have information they didn't want to know)?

Despite the potential ethical issues, these techniques have been successfully used in police investigations. Some of the earliest adoptions of these methods were in the United States where a collaboration between Mark Shriver of Penn State University and DNAprint Genomics led to ancestry testing being marketed and used for forensic casework. One well-publicised use was in Louisiana in 2003 where the techniques were successfully employed to help apprehend a serial rapist and murderer [257]. Early eyewitness testimony had suggested that a white male had been seen fleeing the scene, and this led to the police instigating a futile mass screen of over 1000 white men from the local area, however the ancestry DNA profiling was able to suggest the suspect's genetic makeup was 85% African, and subsequent to this change in investigative direction Derrick Todd Lee was arrested 2 months later, convicted and is currently on death row [258]. Nearer to home, these tests were also employed by the Metropolitan Police in Operation Minstead which investigated the 'Night Stalker' rapist in South London who has been linked to as many as 600 attacks on elderly victims [259]. In this case the tests correctly predicted an individual of Afro-Caribbean origin, but over-reached in the level of detail they could give by specifying that both parents of this man were likely to have come from the Windward Isles (a chain of Islands including St Lucia, Barbados and Trinidad, but not Jamaica where the perpetrator, Delroy Grant, was born) [260]. From the mitochondrial results presented here in section 4.1 it can be seen that there is some differentiation between individuals from Barbados and Jamaica, and indeed when testing these samples with another set of 46 AIMs [261] we get a 93% prediction success (data not shown) for assigning an unknown individual to one of these two islands, however giving predictions this precise is bound to increase the chance of error and evidently the tests and prediction algorithm used by the laboratory reporting this case were not accurate enough and the extra specificity misdirected the police even if the general prediction was correct.

A variation on the 34-plex SNP prediction system has also been used by our Spanish collaborators in Santiago de Compostella during the Madrid train bombing investigation [262]. In March 2004, multiple improvised explosive devices (IEDs)

were placed aboard trains in the Madrid commuter area killing 191 people and injuring a further 1,755. Investigation into the terrorist act involved processing numerous DNA samples, in many cases low level trace samples, from varied items including the bomb making locations and the one IED that failed to detonate. Seven samples produced profiles that didn't match the suspects, and it was hoped at least one of these would lead to the still unknown ringleader. Ancestry testing of these samples using the jointly developed 34-plex and training sets from Spanish and Moroccan populations allowed the assignment of either Spanish or North African ancestry to the individual leaving the trace on the particular item and directed investigators in their search. This was especially tricky considering the extremely close geographical distance between Southern Spain and North Africa and the associated gene flow between the two populations. Three of the samples gave profiles very strongly predicted to be North African, including a toothbrush that was later linked to an Algerian man who has fled and is the on-going subject of an international arrest warrant.

The Y chromosome classification system we developed in section 3.5 has also been the subject of live casework. Validated and implemented by the specialist DNA unit at the Forensic Science Service in Birmingham, along with a selection of Y-SNPs, this prediction matrix has been used on numerous unsolved cases (personal communication A. Revoir, FSS). In addition, this Y-STR classification system devised at our laboratory and detailed here is also cited in Interpol's Handbook on DNA Data Exchange and Practice [263].

7 Conclusions

Throughout the course of this research, sample sets comprising of individuals from the three principal ethnic groups present within the UK (as determined by the census) have been extensively tested across a range of different genetic markers. This has allowed characterisation of these populations, both in terms of genetic composition and substructure, which has a multitude of benefits, both from a population genetics and forensic perspective.

A robust, sensitive and reproducible set of PCR reactions were developed and used to profile 750 individuals from the three main UK populations (Caucasian, Afro-Caribbean and South Asian) for 11 Y-STR markers. The resulting databases were not only published but by providing haplotype frequency estimates this also facilitated the use in the UK of this Y-STR technique in numerous relationship cases, as well as a series of rape and murder cases where analysis was requested from both the prosecution and defence. Characterisation of the Y-STR marker genetics was necessary to assist in interpretation of Y-STR data for many applications, and the publication of this work has been widely cited in further research (a list of all publications arising from research undertaken for this thesis is available in Appendix IV).

A total of 537 individual's mitochondrial DNA was analysed, including 35 for whom the entire 16,569 base pair genome was sequenced. Now that the mitochondrial data has all been collated, this will be forwarded to EMPOP, the first set having already been submitted. The sequence data will also be published, both providing an insight into the mitochondrial genetics of these populations, and providing a basis for comparison either for sequence frequency or geographical location. Further, the full mitochondrial sequencing data will help to refine the known worldwide mitochondrial phylogenetic tree, especially in the Asian branches where there are significant knowledge gaps.

Analysis of the relationship between both the Y-STR haplotypes and mitochondrial sequences with other areas of the globe provides some interesting insights into

population history. The Y chromosome distribution within Ireland is particularly interesting as the majority of samples cluster close together while a subset of 26/155 appear to have very different ancestry: separated from the cluster by 4 mutational steps in two slowly evolving markers, and unusually containing an abundance of unique haplotypes compared with the rest of the world (as sampled on the YHRD) this possibly represents a particular remote Irish genetic signature that split from the remainder of the Irish population many years ago.

In comparison to the UK Caucasian Y-STR data, the Irish Caucasian population overall is shown to be less diverse, but with a high incidence of haplotypes that are otherwise found rarely in the rest of the world, suggesting some degree of conserved patrilineage and genetic independence within a smaller geographic area. This is also mirrored to some extent in the mitochondrial data.

When compared to other European populations, the Y-STR haplotypes from both the Irish and British Caucasian individuals clustered with other populations towards the west of Europe, with far more genetic variation explained by longitude stratification than in the latitude dimension. Comparison of the British Caucasian data with the other two main ethnic groups within Britain: the Afro-Caribbean and South Asian populations, produced results showing a high level of genetic difference between them, with the Afro-Caribbean and Caucasian populations having the least similar haplotype distributions, and the Caucasian and South Asian populations being the most closely related. Phylogenetic network diagrams of the haplotypes showed evidence of some admixture between the populations, although there was appreciable separation between the 3 population groups for most samples.

The mitochondrial sequences from the UK Caucasian population mainly consisted of typically Western European haplogroups, however there were some more unusual sequences possibly providing echoes back to the Viking invasion. The population showed very little maternal admixture (there was one typically African sequence and one typically South Asian sequence), however there were a couple of very rare sequences present that have only previously been seen in Siberia or Uzbekistan and the presence of these in the UK population is unexplained.

The maternal admixture in the UK Afro-Caribbean population is also minimal, with only 5 out of 135 sequences not belonging to the African haplogroup L. This is low compared to the admixture proportions in African-American populations and reflects the different demographic histories. When comparing the modern UK Afro-Caribbean population to two Caribbean islands, the distribution of haplogroups in Jamaica more closely mirrors those seen in the UK while the Barbadian distribution is slightly more distinct – reflecting both the differing histories of the two Caribbean islands and the migration from Jamaica to the UK following the Second World War.

The UK South Asian population is shown to have a diverse set of mitochondrial haplogroups, including some more commonly found in Western Europe. This sub-structuring within the South Asian population is also highlighted in the population specific SNP results where the addition of two extra SNPs taking the marker set from 34 to 36 is able to highlight increased structure within the South Asian data set. This substructure is not surprising considering the diversity in religion, culture, language and ancestral geographical location within the South Asian population.

A set of autosomal SNP markers has been collaboratively developed and tested that can provide excellent predictive success when used to predict continental origin of an unknown DNA sample coming from Africa, Western Europe or East Asia. The analysis of a set of candidate SNPs expected to show Caucasian-South Asian divergence proved very useful in extending the scope of the original population specific SNP marker set from a three population classification system to a four population classification system (Africa, Western Europe, East Asia, and South Asian). When using the stringent one-out cross-validation assessment of classification success, the SNP set has a 100% correct prediction rate for the three-population system and a rate of over 94% for the four-population system – using the 'apparent success' assessment method this rate climbs to 99.2%. A prediction system based on mitochondrial haplogroup membership correctly classified 90% of samples from the three major UK populations, while a decision tree based on Y-STR haplotype managed an 80% classification success rate for the same populations. Despite the lower predictive success, the Y-STR system does have the additional benefit that profile generation is now relatively easy (certainly compared to the other two methods) due to the subsequent development of commercial kits, and the fact that

profiles will already be available in some criminal investigations due to normal DNA testing procedures.

Further work is on-going to produce more specific data for particular African and South Asian populations to see if this aids the classification –generation of Somalian data is expected to be especially useful. The mitochondrial sequencing of additional South Asian individuals would also help elucidate which sub-lineages are more prevalent in the sub-continent as opposed to Europe, and also help to fill out the missing branches in areas of the mitochondrial phylogenetic tree, especially haplogroup R.

Looking at the success rates, it is clear that the SNP predictor system achieves the best results, and the addition of a qualitative measure of confidence in the prediction means that this would be recommended as the primary method of choice for population of origin testing. The addition of a 'not determined' classification would increase the success rate of the Y-STR test to 84% and the SNP test to 100% (even when including the 20 problematic samples re-checked with the SNPs) and would only reduce the number of cases in which these tests were applicable to 90% and 91% respectively. By combining all 3 methods together (or only 2 if the sample is female), it has been demonstrated that an even more secure prediction can be achieved. None of the misclassifying samples tested with the SNPs and additional systems would have been reported wrongly, although some of them would have been classified as 'not determined'. A particularly useful application of combining more than one method is in cases where a weak SNP prediction is obtained. Here the specific mitochondrial haplogroup, or the worldwide location of samples with identical Y-STR types *via* the YHRD, might give a good indication that the tested sample doesn't fit well within any of the four main UK population groups, but instead belongs either to the periphery of one of those groups or ancestrally matches with another population entirely.

Overall, the aim of ancestry prediction from an unknown DNA sample within the UK is shown to be a largely achievable goal. The three classification methods described are shown to have high success rates, much more so than the previous reported attempt within the UK, and could provide crucial intelligence leads in criminal cases

with no suspects, as has already been shown. With the development of other genetic tests to discern physical traits from a DNA sample, e.g. eye colour, it seems clear that population of origin prediction is only the most basic of information that may soon be available to investigative officials.

Appendix I – Autosomal STR Data

Autosomal STR data for the families presenting with Y-STR mutations. Paternity indexes calculated from this data are presented in Table 3.1. Two families needed additional testing, one because the relationship was that of siblings rather than parent-child, and one because an autosomal mutation was also detected, and therefore a whole battery of extra tests were necessary to confirm paternity.

Mutated Marker	Mutation	Relationship	D3S1358		THO1		D21S11		D18S51		Penta E		D5S818		D13S317		D7S820		D16S539		CSF1PO		Penta D		VWA		D8S1179		TPOX		FGA	
DYS385	13-14	Father	15	16	9.3		27	29	13	15	13		11	12	11	12	8	9	10	12	10		13	14	17	18	9	14	8		19	20
		Mother	15	17	7	9	29	31.2	15		9	14	11	12	11	13	8	11	9	12	10	12	9	10	15	18	12	15	8		23	24
	13-15	Child	15	17	7	9.3	29	31.2	13	15	9	13	11	12	12	13	8	11	9	10	10		9	14	17	18	9	12	8		20	24
DYS385	11-14	Father	16		9.3		28	31	14	15	7	18	12		11	12	8	9	10	12	11	12	10	13	14	18	14	15	8		18	23
		Mother	16	17	9.3		31.2	33.1	20		5	19	11		13	14	10	11	8	12	11	12	5	13	15	17	10	16	8		22	27
	10-14	Child	16	17	9.3		31	33.1	14	20	5	7	11	12	12	14	8	11	8	10	11	12	5	13	15	18	10	14	8		22	23
DYS385	15,17	Father	15	17	7		29	31.2	12	17	13		13		12		10	11	11		8	10	2.2	10	16		13	14	7	9	24	27
		Mother	16	17	7		28	29.0	15	16	11	13	10	12	11	12	8	10	9		12		2.2	12	14	17	14	15	6	9	23	
	15,18	Child	17		7		29	31.2	12	15	11	13	12	13	12		10		9	11	8	12	2.2		14	16	14		6	7	23	27
DYS385	11-15	Brother	16	17	9.3		30	32.2	14	15	8	15	12	13	8	11	10		10	11	12		12	14	14	16	11		8	12	20	24
	12-15	Brother	15	17	9.3		29	30	15	20	13	15	12		8	11	10		10	11	12		7	12	16	18	11	13	11	12	20	24
DYS389II	29	Father	14	16	7	9	29	30	12	13	12	13	10	13	11		12		11		12	13	9	11	14	17	9	13	6	8	19	26
		Mother	15		9	9.3	30.2	31.2	12	18	8	11	11		11		10	11	10	11	10		9		16	21	13	15	8	12	21	22
	30	Child	14	15	9		29	30.2	13	18	8	13	11	13	11		11	12	10	11	10	12	9	11	16	17	9	13	8		22	26
DYS389II	29	Father	15	17	7	9	29	32.2	15	20	14	17	10	12	8	11	8	10	12		11		12	13	15	18	13	14	8	11	20	22
		Mother	15	16	7	9.3	29	31.2	12	17	5	11	11	12	12	14	7	10	12	13	12		10		14	18	11	13	9	11	20	23
	30	Child	15	17	7		29		15	17	11	18	12		8	12	7	10	12	13	11	12	10	12	18		11	13	8	9	20	22
DYS391	11	Father	16	18	8	9.3	30		14	15	5	15	11	12	9	10	9	12	11		11	12	11	14	16	17	10	11	9	11	19	22
		Mother	14	15	6	8	28	31.2	14	16	7		12		8		9	11	11	13	10	11	10	12	14	18	8	15	8	9	20	24
	10	Child	15	18	6	9.3	30	31.2	15	16	5	7	11	12	8	9	9	12	11		10	11	12	14	14	16	8	10	9		22	24
DYS391	10	Father	14	18	8	9	29	31.2	14	16	7	13	12	13	11		11		10	13	12		9	13	17		14		8		22	23
		Mother	14	17	9	9.3	28	31.2	15		11	13	12	13	11	12	9	11	12	13	11		11		14		9	13	8		20	21
	11	Child	14	18	9		29	31.2	15	16	7	13	13		11		9	11	12	13	11	12	9	11	14	17	9	14	8		21	22
DYS19	16	Father	16		7	8	28	29	14	17	10	19	11	12	12	13	9	12	9	10	10	12	11	13	17	18	12	14	8	11	18	23
		Mother	15	16	6	9	29	31.2	10	17	7	10	11	12	11		11		12		11		10	11	18		13	14	8		19	20
	15	Child	16		6	8	29	31.2	10	17	7	10	11	12	11	13	9	11	9	12	11	12	11	13	18		12	13	8	11	20	23
DYS389I	15	Father	15	18	7	9	31	32.2	14		11	17	11		10	11	8	11	13	14	11	12	10	11	15	17	14	15	8	11	21	25
		Mother	15	17	9.3		30		12	18	7	14	12	13	11	12	10	11	12		12	13	12		19		10	13	8	11	22	23
	14	Child	15		7	9.3	30	31	14	18	11	14	11	12	11	12	8	11	12	14	11	13	10	12	17	19	13	14	11		21	22
DYS437	16	Father	15	17	9.3		30	31.2	14	16	7	12	11	12	8	12	12	14	11	12	10	11	12	13	16	17	13	14	8	11	21	23
		Mother	13	16	9.3		29	32.0	14	16	5	7	11		9	12	8	12	9	13	10	11	13		14		15		8		20	24
	17	Child	16	17	9.3		31.2	32	14	16	5	7	11	12	12		12	14	9	12	10		13		14	17	14	15	8	11	23	24
DYS439	11	Father	16	17	9.3		31	32.2	13	17	7	12	12	13	8	12	8	13	8	11	12		10	11	17	19	10	12	8		22	24
		Mother	16		9		29	33.2	12	13	7	11	11	13	10		10	11	8	13	11		9	11	16	20	9		8		20	
	10	Child	16	17	9	9.3	31	33.2	12	13	11	12	12	13	10	12	8	11	8	13	11	12	9	10	16	17	9	12	8		20	24
DYS460	11	Father	14	18	6	7	29	31.2	16	17	11	14	12	13	12		10		8	10	10	12	2.2	12	14	19	12	14	8	11	24	
		Mother	15	17	7	8	27		19	21	8	13	12		11	12	8		9		10	12	7	11	16		14	15	8		27	44.2
	10	Child	14	17	7	8	27	31.2	16	21	8	11	12	13	11	12	8	10	8	9	10	12	2.2	7	16	19	12	14	8		24	27

Mutated Marker	Mutation	Relationship	D2S1338	D19S433	LPL	F13B	FES/FPS	F13A01	D12S391	SE33
DYS385	13-14	Father								
		Mother								
	13-15	Child								
DYS385	11-14	Father								
		Mother								
	10-14	Child								
DYS385	15,17	Father								
		Mother								
	15,18	Child								
DYS385	11-15	Brother	20 24	15 17						
	12-15	Brother	20 23	15 17						
DYS389II	29	Father								
		Mother								
	30	Child								
DYS389II	29	Father	16 24	13 16	12 13	9	10	7	20	22.2 25.2
		Mother	20 24	12 14	9 11	6 8	11 12	3.2 4	26	17 31.2
	30	Child	20 24	12 13	9 12	6 9	10 11	3.2 7	20 26	17 22.2
DYS391	11	Father								
		Mother								
	10	Child								
DYS391	10	Father								
		Mother								
	11	Child								
DYS19	16	Father								
		Mother								
	15	Child								
DYS389I	15	Father								
		Mother								
	14	Child								
DYS437	16	Father								
		Mother								
	17	Child								
DYS439	11	Father								
		Mother								
	10	Child								
DYS460	11	Father								
		Mother								
	10	Child								

Appendix II – Ballard *et al.* 2005, FSI (152) 289-305

Appendix III - Ballard *et al.* 2006, FSI (161) 64-68

Appendix IV – List of associated publications

D. Ballard, E. Musgrave-Brown, A. Salas, C. Thacker, D. Syndercombe Court, Mitochondrial analysis of a British Afro-Caribbean population, Progress In Forensic Genetics 10 (2004) 389-391.

D. J. Ballard, C. Phillips, C.R. Thacker, C. Robson, A.P. Revoir, D. Syndercombe Court, Y chromosome STR haplotypes in three UK populations. Forensic Science International, 2005. 152(2-3): p. 289-305

D.J. Ballard, C. Phillips, G. Wright, C.R. Thacker, C. Robson, A.P. Revoir, D. Syndercombe Court, A study of mutation rates and the characterisation of intermediate, null and duplicated alleles for 13 Y chromosome STRs. Forensic Science International, 2005. 155(1): p. 65-70.

D.J. Ballard, C. Phillips, C.R. Thacker, D. Syndercombe Court, Y chromosome STR haplotype data for an Irish population. Forensic Science International, 2006. 161(1): p. 64-8.

D. Syndercombe Court, **D. Ballard**, C. Phillips, A. Revoir, C. Robson, C. Thacker, Comparison of Y-chromosome haplotypes in three racial groups and the possibility of predicting ethnic origin, Progress In Forensic Genetics 9 (2003) 67-69.

L. Roewer, PJP Croucher, S Willuweit, TT Lu, M Kayser, R Lessig, P de Knijff, MA Jobling, C Tyler-Smith, M Krawczak, **Forensic Y Chromosome Research Group** Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Human Genetics (2005) 116:279-291.

C. Phillips, J. Sanchez, M. Fontadevila, A. Gómez-Tato, J. Alvarez-Dios, M. Calaza, M. Casares de Cal, **D. Ballard,** A. Salas, A. Carracedo, A compact population analysis test using 32 SNPs with highly diverse allele frequency distributions, Progress In Forensic Genetics 11 (2006) 58-60.

K.A. Harris, C.R. Thacker, **D. Ballard**, C. Harrison, E. Musgrave-Brown, D. Syndercombe Court, An investigation into the genetic structure of a Barbadian population, Progress In Forensic Genetics 11 (2006) 412-414.

C. Phillips, A. Salas, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, **D. Ballard**, M.V. Lareu, Á. Carracedo, The SNPforID Consortium, Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs, Forensic Science International: Genetics - December 2007 (Vol. 1, Issue 3, Pages 273-280, DOI: 10.1016/j.fsigen.2007.06.008)

C.D. Harrison, **D.J. Ballard**, J. Patel, E. Musgrave Brown, C. Phillips, C.R. Thacker, Y.D. Syndercombe Court, the SNPforID Consortium , Differentiating European and South Asian individuals using SNPs and pyrosequencing technology, Forensic Science International: Genetics Supplement Series - August 2008 (Vol. 1, Issue 1, Pages 476-478, DOI: 10.1016/j.fsigss.2007.10.192)

Appendix V – Example Y-STR File for Network Analysis

Line 1 lists the Y-STR markers in the order they are included within each haplotype. All distinct haplotypes are then listed below in groups of 3 lines: line 1 includes the haplotype name and population separated by a semicolon while line 2 list the Y-STR alleles contained within the haplotype and line 3 details how many times the haplotype was observed within the population. Specific punctuation and formatting are crucial to enable the text file (saved with a .ych extension) to be imported into the Network software, and hence the formatting is displayed within this appendix with · indicating a space and ¶ indicating a carriage return.

19,389a,389b,390,391,392,393,437,438,439¶
¶
¶
1;Irish;;;¶
14,13,16,25,11,14,13,15,12,12¶
· · · · · · · · · · 6¶
4;Irish;;;¶
14,13,16,24,11,13,13,14,12,10¶
· · · · · · · · · · 1¶
7;Irish;;;¶
14,13,16,25,10,14,13,15,12,13¶
· · · · · · · · · · 1¶
10;Irish;;;¶
14,13,16,23,11,14,13,15,12,13¶
· · · · · · · · · · 1¶
14;Irish;;;¶
14,13,16,25,11,14,13,15,12,13¶
· · · · · · · · · · 1¶
17;Irish;;;¶
14,13,16,25,11,15,13,15,12,13¶
· · · · · · · · · · 1¶
20;Irish;;;¶
14,13,17,24,11,13,13,15,12,12¶
· · · · · · · · · · 1¶
23;Irish;;;¶
15,12,16,22,10,9,13,16,10,12¶
· · · · · · · · · · 1¶
26;Irish;;;¶
14,12,17,24,10,13,13,14,12,12¶
· · · · · · · · · · 1¶
29;Irish;;;¶
14,13,15,23,10,13,13,15,12,13¶
· · · · · · · · · · 1¶
32;Irish;;;¶
14,13,16,23,10,13,14,15,12,11¶
· · · · · · · · · · 1¶
35;Irish;;;¶
14,13,16,23,11,13,13,15,12,12¶
· · · · · · · · · · 2¶
38;Irish;;;¶
14,13,16,24,10,13,13,15,12,12¶
· · · · · · · · · · 5¶
41;Irish;;;¶
14,13,16,24,11,13,13,14,12,11¶
· · · · · · · · · · 2¶
44;Irish;;;¶
14,13,16,24,11,13,13,15,12,12¶
· · · · · · · · · · 9¶
47;Irish;;;¶
14,13,16,24,12,13,13,15,12,12¶
· · · · · · · · · · 1¶
50;Irish;;;¶
14,13,17,24,11,14,13,15,12,12¶
· · · · · · · · · · 1¶
53;Irish;;;¶
14,14,16,24,10,13,13,15,12,13¶
· · · · · · · · · · 1¶

Appendix VI – Example Mitochondrial Sequencing File for Network Analysis

Enclosed are the first two pages of the text file containing the British Asian mitochondrial sequencing data for phylogenetic network analysis. The format of this text file is described in detail in section 2.8.5. A dash within an individual sequence indicates that the base in question has been deleted. Specific punctuation and formatting are crucial to enable the text file (saved with a .rdf extension) to be imported into the Network software, and hence the formatting is displayed within this appendix with · indicating a space and ¶ indicating a carriage return.

· ·;1.0¶
16037;16038;16039;16048;16051;16063;16067;16069;16071;16075;16080
;16086;16092;16093;16095;16108;16111;16114;16124;16126;16129;1613
4;16136;16140;16145;16146;16147;16148;16150;16153;16154;16157;161
58;16162;16163;16164;16165;16166;16167;16168;16169;16169.1;16172;
16173;16174;16176;16177;16178;16179;16180;16182;16183;16184;16185
;16186;16187;16188;16189;16192;16193;16194;16195;16203;16206;1620
7;16209;16213;16214;16215;16217;16218;16219;16220;16221;16222;162
23;16224;16227;16230;16231;16233;16234;16235;16239;16241;16242;16
243;16245;16247;16248;16249;16255;16256;16257;16258;16260;16261;1
6263;16264;16265;16266;16270;16271;16274;16275;16278;16286;16287;
16288;16289;16290;16291;16292;16293;16294;16295;16296;16297;16298
;16299;16300;16301;16302;16304;16305;16309;16311;16312;16316;1631
8;16319;16320;16325;16327;16342;16343;16344;16348;16350;16352;163
54;16355;16356;16357;16359;16360;16362;16365;16368;16390;16381;16
391;16398;16399;16400;16445;16456;16463;16468;16482;16497;16502;1
6519;16524;16526;16527;16544;8;10;44.1;52;55;55.1;57;59;60;64;65.
1;66;72;73;89;93;94;95;114;119;125;127;131;143;146;150;151;152;15
3;173;182;183;185;186;188;189;194;195;196;198;199;200;203;204;207
;214;217;222;225;227;228;234;235;236;239;241;242;244;245;246;247;
249;250;253;257;260;263;264;272;279;282;285;291;291.1;292.1;294.1
;295;297;299;310;316;317;318;324;325;337;338;339;340;357;372;372.
1;373;379;385;393;418;447;452;455.1;456;458;461;462;463;466;471;4
77;480;481;482;485;489;494;497;498;499;507;508;511;513;534;522;52
3;524.1;524.3;524.5;549;573;573.1;10398;1018;2416;3666;1048;12372
;4216;7028;12705;3010;4769;6776;3992;4336;9698;3915;4793;11440;33
33;14470;13759;10394;8843;9380;3505;4491;1653;1391;14766;13928;51
78;709;9090;8584;12285;13105;3915a;1811;10400;¶
10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10
;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;1
0;10;10;10;10;10;10;5;5;10;10;10;10;10;10;10;10;10;10;10;10;10;10
;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;1
0;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;
10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10
;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;1
0;10;10;10;10;10;10;10;10;10;2;10;10;10;10;10;10;10;10;10;10;10;1
0;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;
10;10;10;10;10;10;10;10;8;10;10;10;10;10;10;10;10;10;10;10;10;10;
10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10
;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;1
0;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;10;
10;10;10;10;10;10;2;2;5;5;5;10;10;5;100;90;70;70;40;90;70;100;90;
90;90;90;90;90;70;90;90;90;90;90;90;90;90;90;90;90;90;90;90;80;60
;60;60;80;80;80;70;80;80;¶
>AB86;1;;;;A11;A;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCTTAATACACACTCACTGCCGCCTCACCTGAACC
CTATCCCCCCTTAACATAATAAAACTCTACCATCCTTTCTCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTCCCATCAGCAACTTCTAGTGATCGAGAGTTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTTCCCGTACGCC----CC-
AGTGCGTTTGATCTTGAGATGCTCAGTTTGCGTGTAGAC¶
>A56;1;;;;A4;A;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCTTAATACACACTCACTGCCACCTCACCTGAACC
CTATCCACCCTTAACATAATAAAACTCTACCATCCTTTCCCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTCCCATCAGCAACTTCTAGTGATCGAGAGTTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTTCCCGTACGCCA---CC-
AGTGCGTTTGATCTTGAGATGCTCAGTTTGCGTGTAGAC¶

>AB158;1;;;;B6a;B6a;Asian;;¶
AAGGATCCCTATTTCCCCTTACTTAACCCGTTAAAAAACCC-
TCCCATCACCCCCCCCCCATAAATGCATCACCCTAATACACACTCACTGCCACCTCACCTGAACC
CTACCCACCCTTAACATAATAAAGCTCTACCATCCTTTCTCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTTCTATCAACAACTTCTAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTTCCCGTACGCCA---CC-
AGTGCGTTCGATCTTGAGATGCTCAGTTTCCGTGTAGAC¶
>A87;1;;;;B6a;B6a;Asian;;¶
AAGGATCCCTATTCCCCCTTACTTAACCCGTTAAAAAACCC-
TCCCATCACCCCCCCCCCATAAATGCATCACCCTAATACACACTCACTGCCACCTCACCTGAACC
CTACCCACCCTTAACATAATAAAGCTCTACCATCCTTTCTCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTTCTATCAGCAACTTCTAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTTCCCGTACGCCA---CC-
AGTGCGTTCGATCTTGAGATGCTCAGTTTCCGTGTAGAC¶
>AB185;1;;;;D4*;D;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCTTAATACACACTCACTGCCACCTCACCTGAACC
CTACTCACCCTTAACATAATAAAGCTCTACCATCCTTTCCCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTCCTATCAGCAACTTCTAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTCCCCGTACGCCA---CC-
GGTGCGTTTGATCTTGAGATGCTCAGTTTGAGTGTAGAT¶
>AB11;1;;;;D5a2a1;D;Asian;;¶
AAGGATCCCTATCTCCCCTTGCTTGACCCGTTAAAGAATCC-
TCCCATCACCCCCCCCCCCATAAATGCATCACCTTAATACACACTCACTGCCACCTCATCTGAACC
CTACCCACCCTTAACATAATAAAGCTCTACCATCCTTTCCCTGTGGACTGATAATCAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTTCTATCAGCAACCTCCAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTCCCCGTACGCC----CC-
GGTGCGTTTGATCTTGAGATGCTCAGTTTGAGTGTAGAT¶
>AB101;1;;;;G2a1d2;G;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCCTAATACACACTCACTGCCACCTCACCTGAATC
CTACCCACCCTTAACATAATAAAGCTCTACCATCCTTTCCCTGTGGACTGATAATTAGCTGT-
TT-TTTC-
GTGTAGACTTTTGTCCTATCAGCAACTTCTAGTGATCGAGAATTACATTGATCAAGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTCCCCGTACGCCA---CC-
GGTGCGTTTGATCTTGAGATGCTCAGTTTGCATGTAGAT¶
>AS59;1;;;;H*;H;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCCTAATACACACTCACTGCCACCTCACCTGAACC
CTACCCACCCTTAACATAATAAAGCTCTACCATCCTTTCTCTGTGGACTGATAATCAGCTGT-
TT-TTTC-
GTATAGACTTTTGTCCTATCAGCAACTTCTAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCACCACACAT-AAATCCTTCCCCCTTTTCTTTCCCGTACGCCA---CC-
AGTGCGTCCGATCTTGAGATGCTCAGTTCGCGTGTAGAC¶
>6472;1;;;;H*;H;Asian;;¶
AAGGATCCCTATTTCCCCTTGCTTGACCCGTTAAAAAACCC-
TCCCATCAAACCCCCTCCATAAATGCATCACCCTAATACACACTCACTGCCACCTCACCTGAACC
CTACTCACCCTTAACATAATAAAGCTCTACCATCCTTTCTCTGTGGACTGATAATCAGCTGT-
TT-TTTC-
GTATAGACTTTTGTCCTATCAGCAACTTCTAGTGATCGAGAATTACATTGATCAGGCATTCA---
CACTGCTCCACACAT-AAATCCT-CCCCCTTTTCTTTCCCGTACGCC----CC-
AGTGCGTCCGATCTTGAGATGCTCAGTTCGCGTGTAGAC¶

Appendix VII – Example Structure Input File

Enclosed is the first page of the text file for the 34-plex SNP data formatted in the manner necessary for analysis in the Structure software. Listed first are the names of each marker (34 of them in this case) and then follows the genotypes for each sample. Following the sample identifier (the first sample here is labelled 7290), comes the population it belongs to (in the case of sample 7290 this is population 1) and then the genotype data for each of the 34 markers listed in the same order as the markers were listed in the line above. Each allele within the genotype is listed separately (so there are 68 alleles listed for each sample) and the SNP alleles are converted from DNA bases into numbers (so an A allele is converted to a 1, a C to a 2, a G to a 3, a T to a 4 and an N to a -9) – hence the genotype of the first marker for sample 7290 is 2,4 which translates to a CT heterozygous genotype at locus rs1321333 while the second genotype is -9,-9 which means that the genotype of marker rs917118 in sample 7290 could not be determined and was therefore noted as NN.

Specific punctuation and formatting are crucial to enable the text file (saved with a .txt extension) to be imported into the Structure software, and hence the formatting is displayed within this appendix with ➔ indicating a tab and ¶ indicating a carriage return.

rs1321333 → rs917118 → rs1024116 → rs7897550 → rs722098
→ rs10843344 → rs239031 → rs12913832 → rs2040411
→ rs1978806 → rs773658 → rs10141763 → rs182549 → rs1573020
→ rs896788 → rs2065160 → rs2572307 → rs2303798 → rs2065982
→ rs3785181 → rs881929 → rs1498444 → rs1426654 → rs2026721
→ rs4540055 → rs16891982 → rs1335873 → rs1886510
→ rs730570 → rs5030240 → rs2304925 → rs5997008 → rs3827760
→ rs2814778 → → → → → → → → → → → →
→ → → → → → → → → → → → → →
→ → → → ¶
7290 → 1 → 2 → 4 → -9 → -9 → 3 → 3 → 2 → 4 → 3
→ 3 → 2 → 2 → 4 → 4 → 3 → 3 → 1 → 3 → 4 → 4 → 2
→ 2 → 1 → 1 → 4 → 4 → 1 → 1 → 2 → 2 → 1 → 1 → 3
→ 3 → 2 → 2 → 1 → 1 → 2 → 2 → 3 → 3 → 1 → 1 → 4
→ 4 → 3 → 3 → 1 → 1 → 2 → 2 → 1 → 1 → 1 → 3 → 4
→ 4 → 2 → 3 → 4 → 4 → 2 → 2 → 1 → 1 → 4 → 4 ¶
7419 → 1 → 2 → 2 → 3 → -9 → 1 → 1 → 2 → 4 → 1 → 3
→ 2 → 2 → 4 → 4 → 3 → 3 → 1 → 1 → 4 → 4 → 2 → 2
→ 1 → 1 → 4 → 4 → 1 → 1 → 4 → 4 → 1 → 1 → 3 → 3
→ 2 → 2 → 1 → 3 → 2 → 4 → 3 → 3 → 1 → 1 → 4 → 4
→ 3 → 3 → 1 → 1 → 2 → 2 → 1 → 1 → 1 → 3 → 4 → 4
→ 2 → 2 → 3 → 4 → 2 → 2 → 1 → 1 → 4 → 4 ¶
7552 → → 2 → 4 → 3 → -9 → 1 → 3 → 2 → 4 → 1 → 1
→ 2 → 4 → 4 → 4 → 1 → 1 → 1 → 1 → 4 → 4 → 2 → 2
→ 1 → 1 → 2 → 2 → 1 → 1 → 2 → 2 → 1 → 1 → 3 → 3
→ 2 → 2 → 1 → 1 → 2 → 2 → 3 → 4 → 2 → 2 → 4 → 4
→ 1 → 3 → 1 → 1 → 2 → 3 → 1 → 1 → 1 → 1 → 4 → 4
→ 2 → 2 → 4 → 4 → 2 → 2 → 1 → 1 → 4 → 4 ¶
7554 → 1 → 2 → 2 → -9 → -9 → 1 → 3 → 2 → 4 → 1
→ 3 → 2 → 4 → 4 → 4 → 3 → 3 → 3 → 3 → 4 → 4 → 2
→ 2 → 1 → 1 → 4 → 4 → 1 → 1 → 2 → 2 → 1 → 1 → 3
→ 3 → -9 → -9 → 1 → 1 → 2 → 2 → 3 → 4 → 1 → 2
→ 4 → 4 → 3 → 3 → 1 → 1 → 2 → 2 → 1 → 4 → 1 → 3
→ 4 → 4 → 1 → 2 → 4 → 4 → 2 → 2 → 1 → 1 → 4 → 4
¶
8637 → 1 → 2 → 2 → 1 → 3 → 1 → 3 → 2 → 4 → 1 → 1
→ 4 → 4 → 4 → 4 → 3 → 3 → 1 → 1 → 4 → 4 → 2 → 2
→ 1 → 1 → 2 → 4 → 1 → 1 → 2 → 2 → 1 → 1 → 3 → 3
→ 2 → 2 → 1 → 1 → 2 → 2 → 4 → 4 → 1 → 1 → 4 → 4
→ 3 → 3 → 1 → 4 → 2 → 2 → 1 → 4 → 1 → 3 → 2 → 4
→ 3 → 3 → 4 → 4 → 2 → 2 → 1 → 1 → 4 → 4 ¶
10099 → 1 → 2 → 4 → 3 → -9 → 1 → 1 → 2 → 4 → 1
→ 3 → 2 → 2 → 4 → 4 → 1 → 1 → 1 → 3 → 4 → 4 → 2
→ 2 → 1 → 1 → 2 → 2 → 1 → 1 → 2 → 2 → 1 → 1 → 3
→ 3 → 2 → 2 → 1 → 1 → 2 → 4 → 4 → 4 → 1 → 2 → 4
→ 4 → 3 → 3 → 1 → 4 → 2 → 2 → 1 → 4 → 1 → 3 → 4
→ 4 → 2 → 2 → 3 → 4 → 2 → 2 → 1 → 1 → 4 → 4 ¶
14496 → 1 → 4 → 4 → -9 → -9 → 1 → 3 → 2 → 2 → 1
→ 1 → 2 → 2 → 4 → 4 → 1 → 3 → 1 → 1 → 4 → 4 → 2
→ 2 → 1 → 4 → 2 → 4 → 1 → 1 → 2 → 2 → 1 → 1 → 3
→ 3 → 2 → 2 → 1 → 1 → 2 → 2 → 3 → 4 → 1 → 2 → 4
→ 4 → 1 → 3 → 1 → 4 → 2 → 2 → 1 → 4 → 1 → 3 → 4
→ 4 → 2 → 2 → 4 → 4 → 1 → 2 → 1 → 1 → 4 → 4 ¶

8 References

1. Office_for_National_Statistics, *The UK population at the start of the 21st century*, in *Population Trends 122*Winter 2005. p. 7-17.
2. Strong, R.C., *The story of Britain : a people's history*1996, London: Pimlico, 1998. xi, 607 p.
3. Parfitt, S.A., et al., *The earliest record of human activity in northern Europe*. Nature, 2005. **438**(7070): p. 1008-12.
4. Stoneking, M. and H. Soodyall, *Human evolution and the mitochondrial genome*. Curr Opin Genet Dev, 1996. **6**(6): p. 731-6.
5. Templeton, A., *Out of Africa again and again*. Nature, 2002. **416**(6876): p. 45-51.
6. Trinkaus, E., *Early Modern Humans*. Annual Review of Anthropology, 2005. **34**(1): p. 207-30.
7. Powell, T.G.E., *Celtic Origins:A Stage in the Enquiry*. The Journal of the Royal Anthropological Institute of Great Britiain and Ireland, 1948. **78**(1/2): p. 71-79.
8. Office_for_National_Statistics, *Ethnicity and National Identity in England and Wales 2011*, 2012.
9. Simpson, S., *Non-response to the 1991 Census: the effect on ethnic group enumeration*, in *Ethnicity in the 1991 census. Vol. 1, Demographic characteristics of the ethnic minority populations*, D.A. Coleman and J. Salt, Editors. 1996, H.M.S.O.: London. p. Table 3.1.
10. Office_for_National_Statistics, *2011 Census: Key Statistics for England and Wales, March 2011*, 2012.
11. Office_for_National_Statistics, *Fragmented life courses: the changing profile of Britain's ethnic populations*, in *Population Trends 101*Autumn 2000. p. 6-10.
12. Office_for_National_Statistics, *Population estimates by ethnic group*, in *Population Trends 126*Winter 2006. p. 4.
13. Rowe, S.M., S. Miller, and E.J. Sorscher, *Cystic fibrosis*. N Engl J Med, 2005. **352**(19): p. 1992-2001.
14. Interntional Human Genome Sequencing Consortium, *Finishing the euchromatic sequence of the human genome*. Nature, 2004. **431**(7011): p. 931-45.
15. Pang, K.C., M.C. Frith, and J.S. Mattick, *Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function*. Trends in genetics : TIG, 2006. **22**(1): p. 1-5.
16. Ponjavic, J., C.P. Ponting, and G. Lunter, *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs*. Genome research, 2007. **17**(5): p. 556-65.
17. International HapMap Project, *Release #22,* http://www.hapmap.org/abouthapmap.html, 2007.
18. Redon, R., et al., *Global variation in copy number in the human genome*. Nature, 2006. **444**(7118): p. 444-454.
19. Levy, S., et al., *The Diploid Genome Sequence of an Individual Human*. PLoS Biology, 2007. **5**(10): p. e254.

20. Ellegren, H., *Microsatellites: simple sequences with complex evolution*. Nat Rev Genet, 2004. **5**(6): p. 435-45.
21. Applied Biosystems, *AmpFlSTR SGM Plus PCR Amplification Kit User's Manual*, 2006. p. 14:12.
22. Levinson, G. and G.A. Gutman, *Slipped-strand mispairing: a major mechanism for DNA sequence evolution*. Mol Biol Evol, 1987. **4**(3): p. 203-21.
23. Sia, E.A., et al., *Microsatellite instability in yeast: dependence on repeat unit size and DNA mismatch repair genes*. Mol Cell Biol, 1997. **17**(5): p. 2851-8.
24. Xu, X., M. Peng, and Z. Fang, *The direction of microsatellite mutations is dependent upon allele length*. Nat Genet, 2000. **24**(4): p. 396-9.
25. Glenn, T.C., et al., *Allelic diversity in alligator microsatellite loci is negatively correlated with GC content of flanking sequences and evolutionary conservation of PCR amplifiability*. Mol Biol Evol, 1996. **13**(8): p. 1151-4.
26. Brinkmann, B., et al., *Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat*. Am J Hum Genet, 1998. **62**(6): p. 1408-15.
27. Weber, J.L. and C. Wong, *Mutation of human short tandem repeats*. Hum Mol Genet, 1993. **2**(8): p. 1123-8.
28. Abecasis, G.R., et al., *An integrated map of genetic variation from 1,092 human genomes*. Nature, 2012. **491**(7422): p. 56-65.
29. Arndt, P.F., D.A. Petrov, and T. Hwa, *Distinct changes of genomic biases in nucleotide substitution at the time of Mammalian radiation*. Mol Biol Evol, 2003. **20**(11): p. 1887-96.
30. Kumar, S. and S. Subramanian, *Mutation rates in mammalian genomes*. Proc Natl Acad Sci U S A, 2002. **99**(2): p. 803-8.
31. Arndt, P.F., T. Hwa, and D.A. Petrov, *Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects*. J Mol Evol, 2005. **60**(6): p. 748-63.
32. Wright, S., *Evolution in Mendelian Populations*. Genetics, 1931. **16**(2): p. 97-159.
33. Topf, A.L., et al., *Ancient human mtDNA genotypes from England reveal lost variation over the last millennium*. Biol Lett, 2007. **3**(5): p. 550-3.
34. Zerjal, T., et al., *Y-chromosomal insights into the genetic impact of the caste system in India*. Human genetics, 2007. **121**(1): p. 137-44.
35. Seldin, M.F., et al., *European Population Substructure: Clustering of Northern and Southern Populations*. PLoS Genetics, 2006. **2**(9): p. e143.
36. Slatkin, M., *A Population-Genetic Test of Founder Effects and Implications for Ashkenazi Jewish Diseases*. The American Journal of Human Genetics, 2004. **75**(2): p. 282-293.
37. Stoneking, M., *Single nucleotide polymorphisms. From the evolutionary past*. Nature, 2001. **409**(6822): p. 821-2.
38. Sulem, P., et al., *Genetic determinants of hair, eye and skin pigmentation in Europeans*. Nat Genet, 2007. **39**(12): p. 1443-52.
39. Chaplin, G., *Geographic distribution of environmental factors influencing human skin coloration*. Am J Phys Anthropol, 2004. **125**(3): p. 292-302.
40. Norton, H.L., et al., *Genetic evidence for the convergent evolution of light skin in Europeans and East Asians*. Mol Biol Evol, 2007. **24**(3): p. 710-22.

41. Valverde, P., et al., *Variants of the melanocyte-stimulating hormone receptor gene are associated with red hair and fair skin in humans*. Nat Genet, 1995. **11**(3): p. 328-30.
42. Pritchard, J.K., M. Stephens, and P. Donnelly, *Inference of population structure using multilocus genotype data*. Genetics, 2000. **155**(2): p. 945-59.
43. Rosenberg, N.A., et al., *Genetic structure of human populations*. Science, 2002. **298**(5602): p. 2381-5.
44. Excoffier, L. and G. Hamilton, *Comment on "Genetic structure of human populations"*. Science, 2003. **300**(5627): p. 1877; author reply 1877.
45. Bamshad, M.J., et al., *Human population genetic structure and inference of group membership*. Am J Hum Genet, 2003. **72**(3): p. 578-89.
46. Lander, E.S., et al., *Initial sequencing and analysis of the human genome*. Nature, 2001. **409**(6822): p. 860-921.
47. Hinds, D.A., et al., *Whole-genome patterns of common DNA variation in three human populations*. Science, 2005. **307**(5712): p. 1072-9.
48. Lao, O., et al., *Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry*. Am J Hum Genet, 2006. **78**(4): p. 680-90.
49. de Knijff, P., et al., *Chromosome Y microsatellites: population genetic and evolutionary aspects*. International Journal of Legal Medicine, 1997. **110**(3): p. 134-49.
50. Zerjal, T., et al., *Geographical, linguistic, and cultural influences on genetic diversity: Y-chromosomal distribution in Northern European populations*. Mol Biol Evol, 2001. **18**(6): p. 1077-87.
51. Richards, M., et al., *Tracing European founder lineages in the Near Eastern mtDNA pool*. Am J Hum Genet, 2000. **67**(5): p. 1251-76.
52. Salas, A., et al., *The making of the African mtDNA landscape*. American Journal of Human Genetics, 2002. **71**(5): p. 1082-111.
53. Lahn, B.T. and D.C. Page, *Four evolutionary strata on the human X chromosome*. Science, 1999. **286**(5441): p. 964-7.
54. Skaletsky, H., et al., *The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes*. Nature, 2003. **423**(6942): p. 825-37.
55. Bull, J.J. and M.G. Bulmer, *The evolution of XY females in mammals*. Heredity, 1981. **47**(Pt 3): p. 347-65.
56. Berta, P., et al., *Genetic evidence equating SRY and the testis-determining factor*. Nature, 1990. **348**(6300): p. 448-50.
57. Sinclair, A.H., et al., *A gene from the human sex-determining region encodes a protein with homology to a conserved DNA-binding motif*. Nature, 1990. **346**(6281): p. 240-4.
58. Foster, J.W. and J.A. Graves, *An SRY-related sequence on the marsupial X chromosome: implications for the evolution of the mammalian testis-determining gene*. Proc Natl Acad Sci U S A, 1994. **91**(5): p. 1927-31.
59. Willard, H.F., *Tales of the Y chromosome.[comment]*. Nature, 2003. **423**(6942): p. 810-1.
60. Graves, J.A., E. Koina, and N. Sankovic, *How the gene content of human sex chromosomes evolved*. Current opinion in genetics & development, 2006. **16**(3): p. 219-24.
61. Simmler, M.C., et al., *Pseudoautosomal DNA sequences in the pairing region of the human sex chromosomes*. Nature, 1985. **317**(6039): p. 692-7.

62. Freije, D., et al., *Identification of a second pseudoautosomal region near the Xq and Yq telomeres*. Science, 1992. **258**(5089): p. 1784-7.
63. Rozen, S., et al., *Abundant gene conversion between arms of palindromes in human and ape Y chromosomes*. Nature, 2003. **423**(6942): p. 873-6.
64. The Y Chromosome Consortium, *A nomenclature system for the tree of human Y-chromosomal binary haplogroups*. Genome Res, 2002. **12**(2): p. 339-48.
65. Jobling, M.A. and C. Tyler-Smith, *The human Y chromosome: an evolutionary marker comes of age*. Nature Reviews Genetics, 2003. **4**(8): p. 598-612.
66. Xue, Y., et al., *Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree*. Current biology : CB, 2009. **19**(17): p. 1453-7.
67. Brion, M., et al., *Introduction of an single nucleodite polymorphism-based "Major Y-chromosome haplogroup typing kit" suitable for predicting the geographical origin of male lineages*. Electrophoresis, 2005. **26**(23): p. 4411-20.
68. Carvalho-Silva, D.R., et al., *The phylogeography of Brazilian Y-chromosome lineages*. Am J Hum Genet, 2001. **68**(1): p. 281-6.
69. Pereira, L., M.J. Prata, and A. Amorim, *Mismatch distribution analysis of Y-STR haplotypes as a tool for the evaluation of identity-by-state proportions and significance of matches--the European picture*. Forensic Science International, 2002. **130**(2-3): p. 147-55.
70. Kayser, M., et al., *Evaluation of Y-chromosomal STRs: a multicenter study*. International Journal of Legal Medicine, 1997. **110**(3): p. 125-33.
71. Gonzalez-Neira, A., et al., *Distribution of Y-chromosome STR defined haplotypes in Iberia*. Forensic Science International, 2000. **110**(2): p. 117-26.
72. Caglia, A., et al., *Increased forensic efficiency of a STR-based Y-specific haplotype by addition of the highly polymorphic DYS385 locus*. Int J Legal Med, 1998. **111**(3): p. 142-6.
73. Gusmao, L., et al., *Robustness of the Y STRs DYS19, DYS389 I and II, DYS390 and DYS393: optimization of a PCR pentaplex*. Forensic Science International, 1999. **106**(3): p. 163-72.
74. Ayub, Q., et al., *Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information*. Nucleic Acids Research, 2000. **28**(2): p. e8.
75. White, P.S., et al., *New, male-specific microsatellite markers from the human Y chromosome*. Genomics, 1999. **57**(3): p. 433-7.
76. Kayser, M., et al., *A comprehensive survey of human Y-chromosomal microsatellites*. American Journal of Human Genetics, 2004. **74**(6): p. 1183-97.
77. Willuweit, S. and L. Roewer, *Y chromosome haplotype reference database (YHRD): Update*. Forensic Science International: Genetics, 2007. **1**(2): p. 83-87.
78. Chen, X.J. and R.A. Butow, *The organization and inheritance of the mitochondrial genome*. Nat Rev Genet, 2005. **6**(11): p. 815-25.
79. Anderson, S., et al., *Sequence and organization of the human mitochondrial genome*. Nature, 1981. **290**(5806): p. 457-65.
80. Andrews, R.M., et al., *Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA*. Nat Genet, 1999. **23**(2): p. 147.

81. Wallace, D.C., M.D. Brown, and M.T. Lott, *Mitochondrial DNA variation in human evolution and disease*. Gene, 1999. **238**(1): p. 211-30.
82. Wallace, D.C., *Mitochondrial DNA mutations in disease and aging*. Environ Mol Mutagen, 2010. **51**(5): p. 440-50.
83. Wallace, D.C., *A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine*. Annu Rev Genet, 2005. **39**: p. 359-407.
84. Foran, D.R., *Relative degradation of nuclear and mitochondrial DNA: an experimental approach*. J Forensic Sci, 2006. **51**(4): p. 766-70.
85. Margulis, L. and D. Bermudes, *Symbiosis as a mechanism of evolution: status of cell symbiosis theory*. Symbiosis, 1985. **1**: p. 101-24.
86. Gray, M.W., G. Burger, and B.F. Lang, *The origin and early evolution of mitochondria*. Genome Biol, 2001. **2**(6): p. REVIEWS1018.
87. Satoh, M. and T. Kuroiwa, *Organization of multiple nucleoids and DNA molecules in mitochondria of a human cell*. Exp Cell Res, 1991. **196**(1): p. 137-40.
88. Robin, E.D. and R. Wong, *Mitochondrial DNA molecules and virtual number of mitochondria per cell in mammalian cells*. J Cell Physiol, 1988. **136**(3): p. 507-13.
89. Torroni, A., et al., *Intracytoplasmic injection of spermatozoa does not appear to alter the mode of mitochondrial DNA inheritance*. Hum Reprod, 1998. **13**(6): p. 1747-9.
90. Jenuth, J.P., et al., *Random genetic drift in the female germline explains the rapid segregation of mammalian mitochondrial DNA*. Nat Genet, 1996. **14**(2): p. 146-51.
91. Ankel-Simons, F. and J.M. Cummins, *Misconceptions about mitochondria and mammalian fertilization: implications for theories on human evolution*. Proceedings of the National Academy of Sciences of the United States of America, 1996. **93**(24): p. 13859-63.
92. Ramalho-Santos, J., *A sperm's tail: the importance of getting it right*. Human reproduction, 2011. **26**(9): p. 2590-1.
93. Levine, B. and Z. Elazar, *Development. Inheriting maternal mtDNA*. Science, 2011. **334**(6059): p. 1069-70.
94. Nishimura, Y., et al., *Active digestion of sperm mitochondrial DNA in single living sperm revealed by optical tweezers*. Proceedings of the National Academy of Sciences of the United States of America, 2006. **103**(5): p. 1382-7.
95. Hagelberg, E., et al., *Evidence for mitochondrial DNA recombination in a human population of island Melanesia*. Proc Biol Sci, 1999. **266**(1418): p. 485-92.
96. Awadalla, P., A. Eyre-Walker, and J.M. Smith, *Linkage disequilibrium and recombination in hominid mitochondrial DNA*. Science, 1999. **286**(5449): p. 2524-5.
97. Oikawa, H., et al., *The specific mitochondrial DNA polymorphism found in Klinefelter's syndrome*. Biochem Biophys Res Commun, 2002. **297**(2): p. 341-5.
98. Bandelt, H.J., et al., *More evidence for non-maternal inheritance of mitochondrial DNA?* Journal of medical genetics, 2005. **42**(12): p. 957-60.
99. Clayton, D.A., *Replication of animal mitochondrial DNA*. Cell, 1982. **28**(4): p. 693-705.

100. Korr, H., et al., *Mitochondrial DNA synthesis studied autoradiographically in various cell types in vivo*. Braz J Med Biol Res, 1998. **31**(2): p. 289-98.
101. Holt, I.J., H.E. Lorimer, and H.T. Jacobs, *Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA*. Cell, 2000. **100**(5): p. 515-24.
102. Bowmaker, M., et al., *Mammalian mitochondrial DNA replicates bidirectionally from an initiation zone*. J Biol Chem, 2003. **278**(51): p. 50961-9.
103. Brown, T.A., et al., *Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism*. Genes Dev, 2005. **19**(20): p. 2466-76.
104. St-Pierre, J., et al., *Topology of superoxide production from different sites in the mitochondrial electron transport chain*. J Biol Chem, 2002. **277**(47): p. 44784-90.
105. Harman, D., *The biologic clock: the mitochondria?* J Am Geriatr Soc, 1972. **20**(4): p. 145-7.
106. Mandavilli, B.S., J.H. Santos, and B. Van Houten, *Mitochondrial DNA repair and aging*. Mutat Res, 2002. **509**(1-2): p. 127-51.
107. Alaeddini, R., S.J. Walsh, and A. Abbas, *Forensic implications of genetic analyses from degraded DNA--a review*. Forensic Sci Int Genet, 2009. **4**(3): p. 148-57.
108. Teoule, R. and J. Cadet, *Radiation-induced degradation of the base component in DNA and related substances--final products*. Mol Biol Biochem Biophys, 1978. **27**: p. 171-203.
109. Brown, W.M., M. George, Jr., and A.C. Wilson, *Rapid evolution of animal mitochondrial DNA*. Proc Natl Acad Sci U S A, 1979. **76**(4): p. 1967-71.
110. Howell, N., et al., *The pedigree rate of sequence divergence in the human mitochondrial genome: there is a difference between phylogenetic and pedigree rates*. Am J Hum Genet, 2003. **72**(3): p. 659-70.
111. Tamura, K. and M. Nei, *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees*. Mol Biol Evol, 1993. **10**(3): p. 512-26.
112. Coller, H.A., et al., *High frequency of homoplasmic mitochondrial DNA mutations in human tumors can be explained without selection*. Nat Genet, 2001. **28**(2): p. 147-50.
113. Elson, J.L., et al., *Random intracellular drift explains the clonal expansion of mitochondrial DNA mutations with age*. Am J Hum Genet, 2001. **68**(3): p. 802-6.
114. Taylor, R.W., et al., *Mitochondrial DNA mutations in human colonic crypt stem cells*. J Clin Invest, 2003. **112**(9): p. 1351-60.
115. Meissner, C. and S. Ritz-Timme, *Molecular pathology and age estimation*. Forensic Science International, 2010. **203**(1-3): p. 34-43.
116. Cao, L., et al., *New evidence confirms that the mitochondrial bottleneck is generated without reduction of mitochondrial DNA content in early primordial germ cells of mice*. PLoS Genet, 2009. **5**(12): p. e1000756.
117. Howell, N., I. Kubacka, and D.A. Mackey, *How rapidly does the human mitochondrial genome evolve?* Am J Hum Genet, 1996. **59**(3): p. 501-9.
118. Gondos, B., *Comparative studies of normal and neoplastic ovarian germ cells: 2. Ultrastructure and pathogenesis of dysgerminoma*. Int J Gynecol Pathol, 1987. **6**(2): p. 124-31.

119. Cao, L., et al., *The mitochondrial bottleneck occurs without reduction of mtDNA content in female mouse germ cells*. Nat Genet, 2007. **39**(3): p. 386-90.
120. Cree, L.M., et al., *A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes*. Nat Genet, 2008. **40**(2): p. 249-54.
121. Wai, T., D. Teoli, and E.A. Shoubridge, *The mitochondrial DNA genetic bottleneck results from replication of a subpopulation of genomes*. Nat Genet, 2008. **40**(12): p. 1484-8.
122. Koehler, C.M., et al., *Replacement of bovine mitochondrial DNA by a sequence variant within one generation*. Genetics, 1991. **129**(1): p. 247-55.
123. Parsons, T.J., et al., *A high observed substitution rate in the human mitochondrial DNA control region*. Nat Genet, 1997. **15**(4): p. 363-8.
124. Wonnapinij, P., P.F. Chinnery, and D.C. Samuels, *Previous estimates of mitochondrial DNA mutation level variance did not account for sampling error: comparing the mtDNA genetic bottleneck in mice and humans*. Am J Hum Genet, 2010. **86**(4): p. 540-50.
125. Brandstatter, A., et al., *Mitochondrial DNA control region sequences from Nairobi (Kenya): inferring phylogenetic parameters for the establishment of a forensic database*. Int J Legal Med, 2004. **118**(5): p. 294-306.
126. Excoffier, L. and Z. Yang, *Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees*. Mol Biol Evol, 1999. **16**(10): p. 1357-68.
127. Malyarchuk, B.A., et al., *Analysis of phylogenetically reconstructed mutational spectra in human mitochondrial DNA control region*. Hum Genet, 2002. **111**(1): p. 46-53.
128. Hauswirth, W.W., et al., *Heterogeneous mitochondrial DNA D-loop sequences in bovine tissue*. Cell, 1984. **37**(3): p. 1001-7.
129. Butler, J.M. and B.C. Levin, *Forensic applications of mitochondrial DNA*. Trends Biotechnol, 1998. **16**(4): p. 158-62.
130. Lutz-Bonengel, S., et al., *Analysis of mitochondrial length heteroplasmy in monozygous and non-monozygous siblings*. Int J Legal Med, 2008. **122**(4): p. 315-21.
131. Irwin, J.A., et al., *Investigation of heteroplasmy in the human mitochondrial DNA control region: a synthesis of observations from more than 5000 global population samples*. J Mol Evol, 2009. **68**(5): p. 516-27.
132. van Oven, M. and M. Kayser, *Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation*. Hum Mutat, 2009. **30**(2): p. E386-94.
133. McDougall, I., F.H. Brown, and J.G. Fleagle, *Stratigraphic placement and age of modern humans from Kibish, Ethiopia*. Nature, 2005. **433**(7027): p. 733-6.
134. White, T.D., et al., *Pleistocene Homo sapiens from Middle Awash, Ethiopia*. Nature, 2003. **423**(6941): p. 742-7.
135. Tishkoff, S.A. and B.C. Verrelli, *Patterns of human genetic diversity: implications for human evolutionary history and disease*. Annu Rev Genomics Hum Genet, 2003. **4**: p. 293-340.
136. Forster, P. and S. Matsumura, *Evolution. Did early humans go north or south?* Science, 2005. **308**(5724): p. 965-6.

137. Macaulay, V., et al., *Single, rapid coastal settlement of Asia revealed by analysis of complete mitochondrial genomes*. Science, 2005. **308**(5724): p. 1034-6.
138. Lahr, M.M. and R. Foley, *Multiple dispersals and modern human origins*. Evolutionary Anthropology, 1994. **3**(2): p. 48-60.
139. Stringer, C.B. and P. Andrews, *Genetic and fossil evidence for the origin of modern humans*. Science, 1988. **239**(4845): p. 1263-8.
140. Lewis, M.P.e., *Ethnologue: Languages of the World, 16th edition*. 16th Edition ed2009, Dallas, Texas: SIL International.
141. Garrigan, D., et al., *Inferring human population sizes, divergence times and rates of gene flow from mitochondrial, X and Y chromosome resequencing data*. Genetics, 2007. **177**(4): p. 2195-207.
142. Tishkoff, S.A., et al., *The genetic structure and history of Africans and African Americans*. Science, 2009. **324**(5930): p. 1035-44.
143. Campbell, M.C. and S.A. Tishkoff, *The evolution of human genetic and phenotypic variation in Africa*. Curr Biol, 2010. **20**(4): p. R166-73.
144. Behar, D.M., et al., *The dawn of human matrilineal diversity*. Am J Hum Genet, 2008. **82**(5): p. 1130-40.
145. Sun, C., et al., *The dazzling array of basal branches in the mtDNA macrohaplogroup M from India as inferred from complete genomes*. Mol Biol Evol, 2006. **23**(3): p. 683-90.
146. Kivisild, T., et al., *The emerging limbs and twigs of the East Asian mtDNA tree*. Mol Biol Evol, 2002. **19**(10): p. 1737-51.
147. Palanichamy, M.G., et al., *Phylogeny of mitochondrial DNA macrohaplogroup N in India, based on complete sequencing: implications for the peopling of South Asia*. Am J Hum Genet, 2004. **75**(6): p. 966-78.
148. Torroni, A., et al., *Asian affinities and continental radiation of the four founding Native American mtDNAs*. Am J Hum Genet, 1993. **53**(3): p. 563-90.
149. Achilli, A., et al., *The molecular dissection of mtDNA haplogroup H confirms that the Franco-Cantabrian glacial refuge was a major source for the European gene pool*. Am J Hum Genet, 2004. **75**(5): p. 910-8.
150. Brandstatter, A., T.J. Parsons, and W. Parson, *Rapid screening of mtDNA coding region SNPs for the identification of west European Caucasian haplogroups*. Int J Legal Med, 2003. **117**(5): p. 291-8.
151. Pereira, L., et al., *Evaluating the forensic informativeness of mtDNA haplogroup H sub-typing on a Eurasian scale*. Forensic Sci Int, 2006. **159**(1): p. 43-50.
152. Loogvali, E.L., et al., *Disuniting uniformity: a pied cladistic canvas of mtDNA haplogroup H in Eurasia*. Mol Biol Evol, 2004. **21**(11): p. 2012-21.
153. Macaulay, V., et al., *The emerging tree of West Eurasian mtDNAs: a synthesis of control-region sequences and RFLPs*. American Journal of Human Genetics, 1999. **64**(1): p. 232-49.
154. Green, R.E., et al., *A draft sequence of the Neandertal genome*. Science, 2010. **328**(5979): p. 710-22.
155. Walsh, S., et al., *DNA-based eye colour prediction across Europe with the IrisPlex system*. Forensic science international. Genetics, 2012. **6**(3): p. 330-40.
156. Hannum, G., et al., *Genome-wide methylation profiles reveal quantitative views of human aging rates*. Molecular cell, 2013. **49**(2): p. 359-67.

157. Zubakov, D., et al., *Estimating human age from T-cell DNA rearrangements*. Current biology : CB, 2010. **20**(22): p. R970-1.
158. Naik, G., *To Sketch a Thief: Genes Draw Likeness of Suspects*, in *The Wall Street Journal*2009.
159. Kayser, M. and P. de Knijff, *Improving human forensics through advances in genetics, genomics and molecular biology*. Nature reviews. Genetics, 2011. **12**(3): p. 179-92.
160. *The forensic use of bioinformation: ethical issues*, 2007, Nuffield Council on Bioethics.
161. Kayser, M. and P.M. Schneider, *DNA-based prediction of human externally visible characteristics in forensics: motivations, scientific challenges, and ethical considerations*. Forensic science international. Genetics, 2009. **3**(3): p. 154-61.
162. *Human Tissue Act 2004*, 2004, Her Majesty's Stationary Office.
163. Walsh, P.S., D.A. Metzger, and R. Higuchi, *Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material*. Biotechniques, 1991. **10**(4): p. 506-13.
164. *EZ1 DNA Handbook, 2nd Edition*2004: Qiagen.
165. Sanchez, J.J. and P. Endicott, *Developing multiplexed SNP assays with special reference to degraded DNA templates*. Nat Protoc, 2006. **1**(3): p. 1370-8.
166. Vallone, P.M. and J.M. Butler, *AutoDimer: a screening tool for primer-dimer and hairpin structures*. Biotechniques, 2004. **37**(2): p. 226-31.
167. Butler, J.M., *Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (2nd Edition)*2005: Elsevier Academic Press, New York. 688.
168. Applied Biosystems, *BigDye Terminator v3.1 Cycle Sequencing Kit*, 2002.
169. Gusmao, L., et al., *Alternative primers for DYS391 typing: advantages of their application to forensic genetics*. Forensic Sci Int, 2000. **112**(1): p. 49-57.
170. *Technical Manual - PowerPlex 16 System*2007: Promega, P/N TMD012.
171. Underhill, P.A., et al., *Y chromosome sequence variation and the history of human populations.[see comment]*. Nature Genetics, 2000. **26**(3): p. 358-61.
172. Schneider, P.M., et al., *Tandem repeat structure of the duplicated Y-chromosomal STR locus DYS385 and frequency studies in the German and three Asian populations*. Forensic Science International, 1998. **97**(1): p. 61-70.
173. Nei, M., *Molecular Evolutionary Genetics*1987: Columbia University Press, New York. 180.
174. Schneider, S., D. Roessli, and L. Excoffier, *Arlequin: A software for population genetics data analysis. Ver 2.000.* , 2000, Genetics and Biometry Lab, Dept. of Anthropology, University of Geneva.
175. Roewer, L., et al., *Signature of recent historical events in the European Y-chromosomal STR haplotype distribution*. Human Genetics, 2005. **116**(4): p. 279-291.
176. Bandelt, H.J., P. Forster, and A. Rohl, *Median-joining networks for inferring intraspecific phylogenies*. Mol Biol Evol, 1999. **16**(1): p. 37-48.
177. Polzin, T. and S.V. Daneshmand, *On Steiner trees and minimum spanning trees in hypergraphs*. Operations Research Letters, 2003. **31**(1): p. 12-20.
178. Buckleton, J., C.M. Triggs, and S.J. Walsh, *Forensic DNA Evidence Interpretation*2005: CRC Press.
179. Ward, R.H., et al., *Extensive mitochondrial diversity within a single Amerindian tribe*. Proceedings of the National Academy of Sciences of the United States of America, 1991. **88**(19): p. 8720-4.

180. Vigilant, L., et al., *Mitochondrial DNA sequences in single hairs from a southern African population*. Proceedings of the National Academy of Sciences of the United States of America, 1989. **86**(23): p. 9350-4.
181. Rand, S., et al., *The GEDNAP blind trial concept part II. Trends and developments*. Int J Legal Med, 2004. **118**(2): p. 83-9.
182. Bauchet, M., et al., *Measuring European population stratification with microarray genotype data*. American Journal of Human Genetics, 2007. **80**(5): p. 948-56.
183. Yang, N., et al., *Examination of ancestry and ethnic affiliation using highly informative diallelic DNA markers: application to diverse and admixed populations and implications for clinical epidemiology and forensic medicine*. Human Genetics, 2005. **118**(3-4): p. 382-92.
184. Yuasa, I., et al., *Distribution of the F374 allele of the SLC45A2 (MATP) gene and founder-haplotype analysis*. Annals of human genetics, 2006. **70**(Pt 6): p. 802-11.
185. Stacey, S.N., et al., *New common variants affecting susceptibility to basal cell carcinoma*. Nature genetics, 2009. **41**(8): p. 909-14.
186. Phillips, C., et al., *Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs*. Forensic science international. Genetics, 2007. **1**(3-4): p. 273-80.
187. Rosenberg, N.A., *distruct: a program for the graphical display of population structure*. Molecular Ecology Notes, 2004. **4**(1): p. 137-138.
188. Phillips, C., et al., *The recombination landscape around forensic STRs: Accurate measurement of genetic distances between syntenic STR pairs using HapMap high density SNP data*. Forensic science international. Genetics, 2012. **6**(3): p. 354-65.
189. Gill, P., et al., *DNA commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs*. Int J Legal Med, 2001. **114**(6): p. 305-9.
190. Lincoln, P.J., *DNA recommendations--further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems*. Forensic Sci Int, 1997. **87**(3): p. 181-4.
191. Bianchi, N.O., et al., *Characterization of ancestral and derived Y-chromosome haplotypes of New World native populations*. Am J Hum Genet, 1998. **63**(6): p. 1862-71.
192. Budowle, B., et al., *Twelve short tandem repeat loci Y chromosome haplotypes: genetic analysis on populations residing in North America*. Forensic Sci Int, 2005. **150**(1): p. 1-15.
193. Decker, A.E., et al., *Analysis of mutations in father-son pairs with 17 Y-STR loci*. Forensic Science International:Genetics, 2007. ***In Press***.
194. Domingues, P.M., et al., *Sub-Saharan Africa descendents in Rio de Janeiro (Brazil): population and mutational data for 12 Y-STR loci*. Int J Legal Med, 2007. **121**(3): p. 238-41.
195. Dupuy, B.M., et al., *Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci*. Human Mutation, 2004. **23**(2): p. 117-24.
196. Gusmao, L., et al., *Mutation rates at Y chromosome specific microsatellites*. Hum Mutat, 2005. **26**(6): p. 520-8.

197. Heyer, E., et al., *Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees*. Human Molecular Genetics, 1997. **6**(5): p. 799-803.
198. Hohoff, C., et al., *Y-chromosomal microsatellite mutation rates in a population sample from northwestern Germany*. Int J Legal Med, 2007. **121**(5): p. 359-63.
199. Kayser, M., et al., *Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs*. American Journal of Human Genetics, 2000. **66**(5): p. 1580-8.
200. Kurihara, R., et al., *Mutations in 14 Y-STR loci among Japanese father-son haplotypes*. Int J Legal Med, 2004. **118**(3): p. 125-31.
201. Lee, H.Y., et al., *Haplotypes and mutation analysis of 22 Y-chromosomal STRs in Korean father-son pairs*. Int J Legal Med, 2007. **121**(2): p. 128-35.
202. Kayser, M. and A. Sajantila, *Mutations at Y-STR loci: implications for paternity testing and forensic analysis*. Forensic Science International, 2001. **118**(2-3): p. 116-21.
203. Vogt, P.H., et al., *Human Y chromosome azoospermia factors (AZF) mapped to different subregions in Yq11*. Human Molecular Genetics, 1996. **5**(7): p. 933-43.
204. Bosch, E. and M.A. Jobling, *Duplications of the AZFa region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility*. Human Molecular Genetics, 2003. **12**(3): p. 341-7.
205. Zerjal, T., et al., *A genetic landscape reshaped by recent events: Y-chromosomal insights into central Asia*. Am J Hum Genet, 2002. **71**(3): p. 466-82.
206. King, T.E. and M.A. Jobling, *Founders, drift, and infidelity: the relationship between Y chromosome diversity and patrilineal surnames*. Mol Biol Evol, 2009. **26**(5): p. 1093-102.
207. Pattison, J.E., *Is it necessary to assume an apartheid-like social structure in Early Anglo-Saxon England?* Proc Biol Sci, 2008.
208. Sanchez, C., et al., *Haplotype frequencies of 16 Y-chromosome STR loci in the Barcelona metropolitan area population using Y-Filer (TM) kit*. Forensic Science International, 2007. **172**(2-3): p. 211-217.
209. Watson, K. *Slavery and Economy in Barbados*. [cited 2012 5/8/12]; Available from: http://www.bbc.co.uk/history/british/empire_seapower/barbados_01.shtml.
210. Parra, E.J., et al., *Estimating African American admixture proportions by use of population-specific alleles*. Am J Hum Genet, 1998. **63**(6): p. 1839-51.
211. Soares, P., et al., *The archaeogenetics of Europe*. Current biology : CB, 2010. **20**(4): p. R174-83.
212. Gonzalez, A.M., et al., *The mitochondrial lineage U8a reveals a Paleolithic settlement in the Basque country*. BMC genomics, 2006. **7**: p. 124.
213. Quintana-Murci, L., et al., *Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa*. Nature genetics, 1999. **23**(4): p. 437-41.
214. Salas, A., et al., *The African diaspora: mitochondrial DNA and the Atlantic slave trade*. American Journal of Human Genetics, 2004. **74**(3): p. 454-65.

215. Parra, E.J., et al., *Ancestral proportions and admixture dynamics in geographically defined African Americans living in South Carolina*. Am J Phys Anthropol, 2001. **114**(1): p. 18-29.
216. Lind, J.M., et al., *Elevated male European and female African contributions to the genomes of African American individuals*. Hum Genet, 2007. **120**(5): p. 713-22.
217. Kayser, M., et al., *Y chromosome STR haplotypes and the genetic structure of U.S. populations of African, European, and Hispanic ancestry*. Genome research, 2003. **13**(4): p. 624-34.
218. Melchior, L., et al., *Evidence of authentic DNA from Danish Viking Age skeletons untouched by humans for 1,000 years*. PloS one, 2008. **3**(5): p. e2214.
219. Melchior, L., et al., *Genetic diversity among ancient Nordic populations*. PloS one, 2010. **5**(7): p. e11898.
220. Reidla, M., et al., *Origin and diffusion of mtDNA haplogroup X*. American Journal of Human Genetics, 2003. **73**(5): p. 1178-90.
221. Starikovskaya, E.B., et al., *Mitochondrial DNA diversity in indigenous populations of the southern extent of Siberia, and the origins of Native American haplogroups*. Annals of human genetics, 2005. **69**(Pt 1): p. 67-89.
222. Volodko, N.V., et al., *Mitochondrial genome diversity in arctic Siberians, with particular reference to the evolutionary history of Beringia and Pleistocenic peopling of the Americas*. American Journal of Human Genetics, 2008. **82**(5): p. 1084-100.
223. Irwin, J.A., et al., *The mtDNA composition of Uzbekistan: a microcosm of Central Asian patterns*. Int J Legal Med, 2010. **124**(3): p. 195-204.
224. Crubezy, E., et al., *Human evolution in Siberia: from frozen bodies to ancient DNA*. BMC evolutionary biology, 2010. **10**: p. 25.
225. Derenko, M., et al., *Phylogeographic analysis of mitochondrial DNA in northern Asian populations*. American Journal of Human Genetics, 2007. **81**(5): p. 1025-41.
226. Topf, A.L., et al., *Tracing the phylogeography of human populations in Britain based on 4th-11th century mtDNA genotypes*. Mol Biol Evol, 2006. **23**(1): p. 152-61.
227. Richards, M.B., et al., *Phylogeography of mitochondrial DNA in western Europe*. Annals of human genetics, 1998. **62**(Pt 3): p. 241-60.
228. Quintana-Murci, L., et al., *Where west meets east: the complex mtDNA landscape of the southwest and Central Asian corridor*. American Journal of Human Genetics, 2004. **74**(5): p. 827-45.
229. Mishmar, D., et al., *Natural selection shaped regional mtDNA variation in humans*. Proceedings of the National Academy of Sciences of the United States of America, 2003. **100**(1): p. 171-6.
230. Majumder, P.P., *The human genetic history of South Asia*. Current biology : CB, 2010. **20**(4): p. R184-7.
231. Metspalu, M., et al., *Most of the extant mtDNA boundaries in south and southwest Asia were likely shaped during the initial settlement of Eurasia by anatomically modern humans*. BMC genetics, 2004. **5**: p. 26.
232. Kumar, S., et al., *The earliest settlers' antiquity and evolutionary history of Indian populations: evidence from M2 mtDNA lineage*. BMC evolutionary biology, 2008. **8**: p. 230.

233. D'Haene, B., J. Vandesompele, and J. Hellemans, *Accurate and objective copy number profiling using real-time quantitative PCR*. Methods, 2010. **50**(4): p. 262-70.
234. Grosse, I., et al., *Analysis of symbolic sequences using the Jensen-Shannon divergence*. Physical review. E, Statistical, nonlinear, and soft matter physics, 2002. **65**(4 Pt 1): p. 041905.
235. Indian_Genome_Variation_Consortium, *Genetic landscape of the people of India: a canvas for disease gene exploration*. Journal of genetics, 2008. **87**(1): p. 3-20.
236. Bamshad, M.J., et al., *Female gene flow stratifies Hindu castes*. Nature, 1998. **395**(6703): p. 651-2.
237. Basu, A., et al., *Ethnic India: a genomic view, with special reference to peopling and structure*. Genome research, 2003. **13**(10): p. 2277-90.
238. Rosenberg, N.A., et al., *Low levels of genetic divergence across geographically and linguistically diverse populations from India*. PLoS genetics, 2006. **2**(12): p. e215.
239. Harrison, C.D., et al., *Differentiating European and South Asian individuals using SNPs and pyrosequencing technology*. Forensic Science International: Genetics Supplement Series, 2008. **1**(1): p. 476-478.
240. *NCBI - Nucleotide - Homo sapiens POU class 2 homeobox 3 (POU2F3), transcript variant 2, mRNA*. [cited 2012 8/11/2012]; Available from: http://www.ncbi.nlm.nih.gov/nuccore/NM_001244682.1.
241. http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=942793. [cited 2012 16/7/2012].
242. Graf, J., R. Hodgson, and A. van Daal, *Single nucleotide polymorphisms in the MATP gene are associated with normal human pigmentation variation*. Human Mutation, 2005. **25**(3): p. 278-84.
243. Sabeti, P.C., et al., *Genome-wide detection and characterization of positive selection in human populations*. Nature, 2007. **449**(7164): p. 913-8.
244. Stokowski, R.P., et al., *A genomewide association study of skin pigmentation in a South Asian population*. American Journal of Human Genetics, 2007. **81**(6): p. 1119-32.
245. Kosoy, R., et al., *Ancestry informative marker sets for determining continental origin and admixture proportions in common populations in America*. Human Mutation, 2009. **30**(1): p. 69-78.
246. Foreman, L.A. and J.A. Lambert, *Genetic differentiation within and between four UK ethnic groups*. Forensic Science International, 2000. **114**(1): p. 7-20.
247. Curtin, P.D., *The Atlantic Slave Trade : A Census*1969: Madison : University of Wisconsin Press, 1969 (1970).
248. Johnson, D.M., *Black migration in America : A social demographic history*1981, Durham NC: Duke University Press.
249. Kivisild, T., et al., *Deep common ancestry of indian and western-Eurasian mitochondrial DNA lineages*. Current biology : CB, 1999. **9**(22): p. 1331-4.
250. Lowe, A.L., et al., *Inferring ethnic origin by means of an STR profile*. Forensic Science International, 2001. **119**(1): p. 17-22.
251. Cho, M.K. and P. Sankar, *Forensic genetics and ethical, legal and social implications beyond the clinic*. Nature genetics, 2004. **36**(11 Suppl): p. S8-12.
252. Cooper, R.S., J.S. Kaufman, and R. Ward, *Race and genomics*. The New England journal of medicine, 2003. **348**(12): p. 1166-70.

253. M'Charek, A., V. Toom, and B. Prainsack, *Bracketing off population does not advance ethical reflection on EVCs: a reply to Kayser and Schneider*. Forensic science international. Genetics, 2012. **6**(1): p. e16-7; author reply e18-9.
254. Spinney, L., *Eyewitness identification: line-ups on trial*. Nature, 2008. **453**(7194): p. 442-4.
255. *The Innocence Project - Facts on Post-Conviction DNA Exonerations*. [cited 2012 25/10/2012]; Available from: http://www.innocenceproject.org/Content/Facts_on_PostConviction_DNA_Exonerations.php.
256. Kayser, M. and P.M. Schneider, *Reply to "Bracketing off population does not advance ethical reflection on EVCs: A reply to Kayser and Schneider" by A. M'charek, V. Toom, and B. Prainsack*. Forensic science international. Genetics, 2012. **6**(1): p. e18-e19.
257. *Mark Shriver Bio, Department of Anthropolgy, Penn State University*. [cited 2012 29/10/2012]; Available from: http://www.anthro.psu.edu/faculty_staff/shriver.shtml.
258. *Genome Tests Nets Suspected Serial Killer*. [cited 2012 29/10/2012]; Available from: http://www.genomenewsnetwork.org/articles/06_03/serial.shtml.
259. *Night Stalker rapist Delroy Grant jailed for 27 years*. [cited 2012 29/10/2012]; Available from: http://www.bbc.co.uk/news/uk-england-london-12857539.
260. *2,00 DNA tests in hunt for 'night stalker'*. The Independent [cited 2012 29/10/2012]; Available from: http://www.independent.co.uk/news/uk/crime/2000-dna-tests-in-hunt-for-night-stalker-1816140.html.
261. Pereira, R., et al., *Straightforward inference of ancestry and admixture proportions through ancestry-informative insertion deletion multiplexing*. PloS one, 2012. **7**(1): p. e29684.
262. Phillips, C., et al., *Ancestry analysis in the 11-M Madrid bomb attack investigation*. PloS one, 2009. **4**(8): p. e6583.
263. *Interpol Handbook on DNA Data Exchange and Practice, Recommendations from the Interpol DNA Monitoring Expert Group*, 2009, OIPC-Interpol: Lyon. p. 79.